GROUPE D'EXPERTS

INDEPENDANTS DE HAUT NIVEAU SUR L'INTELLIGENCE ARTIFICIELLE

CONSTITUE PAR LA COMMISSION EUROPEENNE EN JUIN 2018

LIGNES DIRECTRICES EN MATIERE D'ETHIQUE POUR UNE IA DIGNE DE CONFIANCE

LIGNES DIRECTRICES EN MATIERE D'ETHIQUE pour UNE IA DIGNE DE CONFIANCE

Groupe d'experts de haut niveau sur l'intelligence artificielle

Le présent document a été rédigé par le groupe d'experts de haut niveau sur l'intelligence artificielle (GEHN IA). Les membres du GEHN IA qui y sont cités soutiennent le cadre général pour une IA digne de confiance présenté dans les présentes lignes directrices, sans approuver nécessairement chacune des affirmations formulées dans le document.

Afin de recueillir des commentaires pratiques, les parties prenantes soumettront à une phase pilote la liste d'évaluation pour une IA digne de confiance présentée au chapitre III du présent document. Une version révisée de cette liste d'évaluation tenant compte des commentaires recueillis au cours de la phase pilote sera présentée à la Commission européenne début 2020.

Le GEHN IA est un groupe d'experts indépendants constitué par la Commission européenne en juin 2018.

Personne de contact Nathalie Smuha – coordinatrice du groupe d'experts de haut niveau sur l'IA
Adresse électronique CNECT-HLG-AI@ec.europa.eu

Commission européenne
B-1049 Bruxelles

Document rendu public le X avril 2019.

Un premier projet de ce document a été publié le 18 décembre 2018 et a fait l'objet d'une consultation ouverte à laquelle plus de 500 contributeurs ont apporté des commentaires. Nous souhaitons remercier explicitement et chaleureusement toutes les personnes ayant fait part de leurs commentaires sur le premier projet de ce document. Ces commentaires ont été pris en compte dans le cadre de l'élaboration de cette version révisée.

De plus amples informations sur le groupe d'experts de haut niveau sur l'intelligence artificielle sont disponibles en ligne (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

TABLE DES MATIERES

RÉSUMÉ

(1) Les présentes lignes directrices visent à promouvoir une IA digne de confiance. Une IA digne de confiance présente les **trois caractéristiques** suivantes, qui devraient être respectées tout au long du cycle de vie du système: a) elle doit être **licite**, en assurant le respect des législations et réglementations applicables; b) elle doit être **éthique**, en assurant l'adhésion à des principes et valeurs éthiques; et c) elle doit être **robuste**, sur le plan tant technique que social car, même avec de bonnes intentions, les systèmes d'IA peuvent causer des préjudices involontaires. Toutes ces caractéristiques sont nécessaires en elles-mêmes, mais elles ne sauraient suffire à la réalisation d'une IA digne de confiance. L'idéal serait que ces trois caractéristiques fonctionnent en harmonie et se chevauchent. Si, dans la pratique, des tensions venaient à apparaître entre ces caractéristiques, la société devrait s'efforcer d'y remédier.

(2) Les présentes lignes directrices établissent un **cadre pour parvenir à la réalisation d'une IA digne de confiance**. Ce cadre ne traite pas explicitement de la première caractéristique d'une IA digne de confiance (IA licite)[1]. Il vise plutôt à proposer des orientations pour encourager et garantir une IA éthique et robuste (les deuxième et troisième caractéristiques). S'adressant à l'ensemble des parties prenantes, les présentes lignes directrices cherchent, en plus de présenter une liste de principes éthiques, à fournir des orientations sur la manière dont ces principes peuvent être mis en œuvre dans des systèmes sociotechniques. Ces orientations se présentent sous la forme de trois niveaux d'abstraction, du plus abstrait, au chapitre I, au plus concret, au chapitre III, et se concluant par des exemples de possibilités et de préoccupations graves soulevées par les systèmes d'IA.

I. Sur la base d'une approche fondée sur les droits fondamentaux, le chapitre I recense les **principes éthiques** et les valeurs correspondantes qu'il convient de respecter lors de la mise au point, du déploiement et de l'utilisation de systèmes d'IA.

Orientations essentielles dérivées du chapitre I:

- ✓ Mettre au point, déployer et utiliser des systèmes d'IA en respectant les principes éthiques suivants: *respect de l'autonomie humaine, prévention de toute atteinte, équité et explicabilité.* Reconnaître et résoudre les tensions potentielles entre ces principes.
- ✓ Accorder une attention particulière aux situations concernant des groupes plus vulnérables tels que les enfants, les personnes handicapées et d'autres groupes historiquement défavorisés ou exposés au risque d'exclusion, et aux situations caractérisées par des asymétries de pouvoir ou d'information, par exemple entre les employeurs et les travailleurs, ou entre les entreprises et les consommateurs.[2]
- ✓ Reconnaître et être conscient que les systèmes d'IA apportent certes des avantages considérables aux individus et à la société, mais qu'ils présentent également certains risques et peuvent avoir des incidences négatives, y compris des incidences pouvant s'avérer difficiles à anticiper, à déterminer ou à mesurer (par exemple des incidences sur la démocratie, l'état de droit et la justice distributive, ou sur l'esprit humain même). Adopter des mesures appropriées pour atténuer ces risques, le cas échéant, d'une manière proportionnée à l'ampleur du risque.

[1] Tous les éléments normatifs du présent document ont pour but de refléter les orientations destinées à réaliser les deuxième et troisième caractéristiques d'une IA digne de confiance (une IA éthique et robuste). Ces éléments ne sont par conséquent pas destinés à fournir des conseils juridiques ou à proposer des orientations en matière de conformité avec la législation applicable, bien qu'il soit reconnu qu'une part importante de ces éléments sont dans une certaine mesure déjà présents dans la législation existante. Voir point 21 et suivants à cet égard.

[2] Voir articles 24 à 27 de la charte des droits fondamentaux de l'Union européenne (charte de l'UE), portant sur les droits de l'enfant et des personnes âgées, l'intégration des personnes handicapées et les droits des travailleurs. Voir également l'article 38 portant sur la protection des consommateurs.

II. S'appuyant sur le chapitre I, le chapitre II fournit des orientations sur la manière dont une IA digne de confiance peut être réalisée, en présentant **sept exigences** que tout système d'IA devrait respecter. Des méthodes tant techniques que non techniques peuvent être utilisées aux fins de leur mise en œuvre.

Orientations essentielles dérivées du chapitre II:

- ✓ Veiller à ce que la mise au point, le déploiement et l'utilisation de systèmes d'IA répondent aux exigences d'une IA digne de confiance: 1) action humaine et contrôle humain, 2) robustesse technique et sécurité, 3) respect de la vie privée et gouvernance des données, 4) transparence, 5) diversité, non-discrimination et équité, 6) bien-être sociétal et environnemental, et 7) responsabilité.
- ✓ Envisager des méthodes techniques et non techniques pour garantir la mise en œuvre de ces exigences.
- ✓ Encourager la recherche et l'innovation en vue de contribuer à l'évaluation des systèmes d'IA et de soutenir la mise en œuvre des exigences; diffuser les résultats et les questions ouvertes au grand public, et veiller à ce qu'une formation dans le domaine l'éthique en matière d'IA soit systématiquement dispensée à la nouvelle génération d'experts.
- ✓ Fournir de façon proactive des informations claires aux parties prenantes sur les capacités et les limites des systèmes d'IA, afin de leur permettre de formuler des attentes réalistes, ainsi que sur la manière dont les exigences sont mises en œuvre. Faire preuve de transparence sur le fait qu'elles interagissent avec un système d'IA.
- ✓ Faciliter la traçabilité et l'auditabilité des systèmes d'IA, en particulier dans les contextes ou situations critiques.
- ✓ Associer les parties prenantes tout au long du cycle de vie du système d'IA. Encourager la formation et l'éducation afin que toutes les parties prenantes soient renseignées sur l'IA digne de confiance et formées dans ce domaine.
- ✓ Savoir qu'il peut exister des tensions fondamentales entre différents principes et exigences. Recenser, évaluer, documenter et communiquer de manière continue ces arbitrages et leurs solutions.

III. Le chapitre III fournit une liste d'évaluation concrète mais non exhaustive pour une IA digne de confiance, qui vise à concrétiser les exigences définies au chapitre II. Cette **liste d'évaluation** devra être adaptée au cas d'utilisation spécifique du système d'IA.[3]

Orientations essentielles dérivées du chapitre III:

- ✓ Adopter une évaluation pour une IA digne de confiance lors de la mise au point, du déploiement ou de l'utilisation de systèmes d'IA, et l'adapter au cas d'utilisation spécifique du système.
- ✓ Garder à l'esprit qu'une liste d'évaluation de cette nature ne sera jamais exhaustive. Il ne suffit pas de cocher des cases pour garantir une IA digne de confiance. Il convient de déterminer des exigences et de les mettre en œuvre, d'évaluer des solutions et de veiller à améliorer les résultats tout au long du cycle de vie du système d'IA, et d'y associer les parties prenantes.

(3) La section finale du document vise à concrétiser certaines des questions abordées dans l'ensemble du cadre, en présentant des exemples de possibilités bénéfiques qu'il convient de mettre en œuvre, et les grandes préoccupations soulevées par les systèmes d'IA qu'il convient d'examiner avec soin.

(4) Si l'objectif des présentes lignes directrices est de proposer des orientations relatives aux applications de l'IA en général, en érigeant une base transversale pour parvenir à une IA digne de confiance, des situations différentes posent des défis différents. Il convient par conséquent d'examiner si, en plus de ce cadre

[3] Conformément au champ d'application du cadre établi au point 2, cette liste d'évaluation ne fournit aucun conseil pour veiller à la conformité juridique (IA licite), mais se limite à proposer des orientations pour réaliser les deuxième et troisième caractéristiques d'une IA digne de confiance (une IA éthique et robuste).

transversal, une approche sectorielle est nécessaire, étant donné la mesure dans laquelle les systèmes d'IA sont spécifiques à leurs contextes.

(5) Les présentes lignes directrices ne visent ni à remplacer toute forme actuelle ou future d'élaboration de politiques ou de réglementations ni à en décourager l'introduction. Il faut les considérer comme un document évolutif qu'il conviendra de réviser et mettre à jour au fil du temps afin d'en maintenir la pertinence, à mesure que la technologie, nos environnements sociaux et nos connaissances évolueront. Le présent document est conçu comme le point de départ de la discussion sur «Une IA digne de confiance pour l'Europe».[4] Au-delà de l'Europe, les présentes lignes directrices visent également à encourager la recherche, la réflexion et la discussion sur un cadre éthique pour les systèmes d'IA au niveau mondial.

[4] Cet idéal est destiné à être appliqué aux systèmes d'IA mis au point, déployés et utilisés dans les États membres de l'Union européenne, ainsi qu'aux systèmes mis au point ou produits ailleurs mais déployés et utilisés au sein de l'UE. Lorsqu'il est fait référence à l'«Europe» dans le présent document, ce sont les États membres de l'Union qui sont visés. Toutefois, les présentes lignes directrices aspirent également à être pertinentes en dehors de l'Union. À cet égard, il convient de noter que la Norvège et la Suisse font partie du plan coordonné dans le domaine de l'IA adopté et publié en décembre 2018 par la Commission et les États membres.

A. INTRODUCTION

(6) Dans ses communications du 25 avril 2018 et du 7 décembre 2018, la Commission européenne (ci-après la «Commission») définit sa vision pour l'intelligence artificielle (IA), qui préconise une «IA éthique, sûre et de pointe réalisée en Europe».[5] La vision de la Commission repose sur trois piliers: i) accroître les investissements publics et privés dans l'IA afin d'intensifier le recours à l'IA, ii) se préparer aux changements socioéconomiques, et iii) garantir un cadre éthique et juridique approprié afin de renforcer les valeurs européennes.

(7) Pour soutenir la mise en œuvre de cette vision, la Commission a mis sur pied le groupe d'experts de haut niveau sur l'intelligence artificielle (GEHN IA), un groupe indépendant chargé d'élaborer deux contributions: 1) des lignes directrices en matière d'éthique, et 2) des recommandations en matière de politique et d'investissement dans le domaine de l'IA.

(8) Le présent document contient les lignes directrices en matière d'éthique dans le domaine de l'IA, qui ont été révisées à la suite de nouvelles délibérations de notre groupe à la lumière des commentaires reçus dans le cadre de la consultation publique relative au projet publié le 18 décembre 2018. Il s'appuie en outre sur les travaux du Groupe européen d'éthique des sciences et des nouvelles technologies[6] et s'inspire d'autres efforts similaires.[7]

(9) Au cours des derniers mois, nos 52 membres se sont réunis, ont discuté et ont interagi, sans déroger à la devise européenne: «Unie dans la diversité». Nous sommes convaincus que l'IA est susceptible de transformer la société de manière significative. L'IA n'est pas une fin en soi, mais plutôt un moyen prometteur d'accroître la prospérité humaine, en renforçant ainsi le bien-être individuel et de la société ainsi que le bien commun, et en étant porteur de progrès et d'innovation. Les systèmes d'IA peuvent notamment contribuer à faciliter la réalisation des objectifs de développement durable des Nations unies, tels que promouvoir l'égalité entre les sexes et lutter contre le changement climatique, rationaliser notre utilisation des ressources naturelles, améliorer notre santé, notre mobilité et nos processus de production, et nous aider à surveiller nos progrès par rapport à des indicateurs de durabilité et de cohésion sociale.

(10) Pour parvenir à ces objectifs, les systèmes d'IA[8] doivent être **centrés sur l'humain**, en s'appuyant sur l'engagement de mettre leur utilisation au service de l'humanité et du bien commun, avec pour objectif d'améliorer le bien-être et la liberté des êtres humains. S'ils offrent de brillantes possibilités, les systèmes d'IA soulèvent également certains risques qui doivent être traités de manière appropriée et proportionnée. Nous sommes à présent face à une occasion unique de façonner leur élaboration. Nous voulons pouvoir nous fier aux environnements sociotechniques auxquels ils sont intégrés, et nous voulons que les concepteurs de systèmes d'IA obtiennent un avantage concurrentiel en intégrant une IA digne de confiance à leurs produits et services. Cet objectif nécessite de chercher à **optimiser les avantages offerts par les systèmes d'IA** tout en veillant **à prévenir et réduire le plus possible les risques qu'ils présentent.**

(11) Dans un contexte d'évolution technologique rapide, nous sommes convaincus qu'il est essentiel que la confiance reste le ciment des sociétés, des communautés, des économies et du développement durable. Nous

---

5 COM(2018) 237 et COM(2018) 795. Il convient de noter que le terme «made in Europe» est employé par la Commission dans sa communication. Le champ d'application des présentes lignes directrices englobe non seulement les systèmes d'IA réalisés en Europe, mais également ceux mis au point ailleurs et qui sont déployés ou utilisés en Europe. Tout au long de ce document, nous nous efforçons donc de promouvoir une IA digne de confiance «pour» l'Europe.

6 Le Groupe européen d'éthique des sciences et des nouvelles technologies (GEE) est un groupe consultatif de la Commission.

7 Voir section 3.3 du document COM(2018) 237.

8 Le glossaire figurant à la fin du présent document fournit une définition des systèmes d'IA aux fins de ce même document. Cette définition est davantage détaillée dans un document spécifique élaboré par le GEHN IA et accompagnant les présentes lignes directrices, intitulé «A definition of AI: Main capabilities and scientific disciplines» (Définition de l'IA: principales capacités et disciplines scientifiques).

avons ainsi fait de **l'IA digne de confiance notre ambition fondatrice**; étant donné que les êtres humains et les communautés ne pourront avoir confiance dans le développement de la technologie et dans ses applications que lorsqu'un cadre clair et exhaustif pour la rendre digne de confiance sera en place.

(12) Il s'agit, de notre point de vue, de la voie que devrait suivre l'Europe pour se positionner comme foyer et leader d'une technologie éthique et de pointe. C'est grâce à une IA digne de confiance que, en tant que citoyens européens, nous pourrons bénéficier de ses avantages d'une manière qui reflète nos valeurs fondamentales que sont le respect des droits de l'homme, la démocratie et l'état de droit.

IA digne de confiance

(13) La fiabilité est une condition préalable pour que les personnes et les sociétés mettent au point, déploient et utilisent des systèmes d'IA. S'ils ne démontrent pas qu'ils sont dignes de confiance, les systèmes d'IA – et les êtres humains qui les conçoivent – pourraient être à l'origine de conséquences indésirables susceptibles de nuire à leur utilisation, ce qui empêcherait la réalisation des avantages sociaux et économiques potentiellement vastes qu'apportent les systèmes d'IA. Pour aider l'Europe à obtenir la réalisation de ces avantages, notre vision consiste à faire de l'éthique un pilier essentiel pour garantir et développer une IA digne de confiance.

(14) La confiance dans la mise au point, le déploiement et l'utilisation de systèmes d'IA concerne non seulement les propriétés intrinsèques de la technologie, mais également les qualités des systèmes sociotechniques impliquant des applications d'IA.[9] De manière analogue à des questions de (perte de) confiance dans l'aviation, l'énergie nucléaire ou la sécurité alimentaire, ce ne sont pas uniquement les composantes des systèmes d'IA qui pourraient ou non susciter la confiance, mais le système dans son contexte global. La quête d'une IA digne de confiance concerne donc non seulement la fiabilité du système d'IA en tant que tel, mais requiert également une approche globale et systémique qui englobe la fiabilité de l'ensemble des acteurs et processus qui composent le contexte sociotechnique du système tout au long de son cycle de vie.

(15) Une IA digne de confiance comporte les **trois éléments** suivants, qui doivent être présents tout au long du cycle de vie du système:

1. elle doit être **licite**, en assurant le respect des législations et réglementations applicables;
2. elle doit être **éthique**, en assurant l'adhésion à des principes et valeurs éthiques, et
3. elle doit être **robuste**, sur le plan tant technique que social car, même avec de bonnes intentions, les systèmes d'IA peuvent causer des préjudices involontaires.

(16) Toutes ces caractéristiques sont nécessaires, mais elles ne sauraient suffire à la réalisation d'une IA digne de confiance[10]. L'idéal serait que ces trois caractéristiques fonctionnent en harmonie et se chevauchent. Toutefois, dans la pratique, des tensions peuvent survenir entre ces éléments (par exemple, dans certains cas, le champ d'application et le contenu de la législation existante pourraient ne pas correspondre à des normes éthiques). Il en va de notre responsabilité individuelle et collective en tant que société de veiller à ce que chacune de ces trois caractéristiques contribue à garantir l'avènement d'une IA digne de confiance.[11]

(17) Une approche digne de confiance est essentielle pour permettre une «compétitivité responsable», en établissant les bases sur lesquelles les personnes concernées par des systèmes d'IA peuvent se fier au caractère licite, éthique et robuste de leur conception, de leur mise au point et de leur utilisation. Les présentes lignes directrices visent à encourager une innovation responsable et durable dans le domaine de l'IA

[9] Ces systèmes se composent d'êtres humains, d'acteurs étatiques, d'entreprises, d'infrastructures, de logiciels, de protocoles, de normes, de gouvernance, de législations existantes, de mécanismes de contrôle, de structures d'incitation, de procédures d'audit, de meilleures pratiques, de documentation, et d'autres éléments.

[10] Cela n'exclut pas le fait que des conditions supplémentaires pourraient être (ou devenir) nécessaires.

[11] Cela signifie également que le législateur ou les décideurs politiques pourraient être amenés à revoir le caractère approprié de la législation en vigueur lorsque celle-ci pourrait ne pas correspondre à des principes éthiques.

en Europe. Elles cherchent à ériger l'éthique en pilier essentiel de la mise au point d'une approche unique de l'IA cherchant à favoriser, renforcer et protéger tant la prospérité individuelle des êtres humains que le bien commun de la société. Nous sommes convaincus que cela permettra à l'Europe de s'imposer comme leader mondial d'une IA de pointe, digne de notre confiance individuelle et collective. Ce n'est que si la fiabilité des systèmes d'IA est garantie que les citoyens européens pourront bénéficier pleinement de ses avantages, forts de la conviction que des mesures sont en place pour les protéger contre les risques potentiels.

(18) Tout comme l'utilisation de systèmes d'IA ne s'arrête pas aux frontières nationales, leurs incidences ne s'y arrêtent pas davantage. Des solutions mondiales sont par conséquent nécessaires face aux possibilités et aux défis mondiaux que présente l'IA. Nous encourageons par conséquent l'ensemble des parties prenantes à travailler à l'élaboration d'un cadre mondial pour une IA digne de confiance, en cherchant un consensus international tout en encourageant et en préservant notre approche fondée sur le respect des droits fondamentaux.

Public et champ d'application

(19) Les présentes lignes directrices sont destinées à l'ensemble des parties prenantes de l'IA qui conçoivent, mettent au point, déploient, mettent en œuvre, utilisent l'IA ou sont soumises à ses incidences, et notamment aux entreprises, aux organisations, aux chercheurs, aux services publics, organismes gouvernementaux, institutions, organisations de la société civile, particuliers, travailleurs et consommateurs. Les parties prenantes résolues à réaliser une IA digne de confiance peuvent librement décider d'utiliser les présentes lignes directrices comme méthode pour concrétiser leur engagement, notamment en ayant recours à la liste d'évaluation pratique du chapitre III dans leurs processus de mise au point et de déploiement de systèmes d'IA. Cette liste d'évaluation peut également compléter, et donc intégrer, les processus d'évaluation existants.

(20) L'objectif des présentes lignes directrices est de proposer des orientations relatives aux applications d'IA en général, en érigeant une base transversale pour parvenir à une IA digne de confiance. Toutefois, **des situations différentes posent des défis différents.** Les systèmes d'IA de recommandation musicale ne soulèvent pas les mêmes préoccupations éthiques que les systèmes d'IA proposant des traitements médicaux essentiels. De même, les systèmes d'IA utilisés dans le contexte des relations d'entreprise à consommateur, d'entreprise à entreprise, d'employeur à employé et de la sphère publique aux citoyens ou, plus généralement, dans différents secteurs ou cas d'utilisation, présentent des possibilités et des défis différents. Les systèmes d'IA étant propres à leur contexte, il est par conséquent reconnu que la mise en œuvre des présentes lignes directrices doit être adaptée à l'application spécifique de l'IA. Il convient en outre d'examiner la mesure dans laquelle une approche sectorielle supplémentaire pourrait être nécessaire pour compléter le cadre transversal plus général proposé dans le présent document.

Afin de mieux comprendre la manière dont ces orientations peuvent être mises en œuvre au niveau transversal, ainsi que les questions qui requièrent une approche sectorielle, nous invitons l'ensemble des parties prenantes à tester la liste d'évaluation pour une IA digne de confiance (chapitre III) concrétisant ce cadre et à nous communiquer leurs observations. Sur la base des commentaires recueillis lors de cette phase pilote, nous réviserons la liste d'évaluation des présentes lignes directrices d'ici le début de 2020. La phase pilote débutera d'ici l'été 2019 et se prolongera jusqu'à la fin de l'année. Toutes les parties prenantes intéressées auront la possibilité de participer en manifestant leur intérêt via l'Alliance européenne pour l'IA.

B. CADRE POUR UNE IA DIGNE DE CONFIANCE

(21) Les présentes lignes directrices définissent un cadre pour parvenir à la mise en œuvre d'une IA digne de confiance fondée sur les droits fondamentaux tels que consacrés dans la charte des droits fondamentaux de l'Union européenne (charte de l'UE), et dans le droit international pertinent en matière de droits de l'homme. Ci-dessous, nous abordons brièvement les trois caractéristiques d'une IA digne de confiance.

IA licite

(22) Les systèmes d'IA ne sont pas mis en œuvre dans un monde sans loi. Un ensemble de règles contraignantes aux niveaux européen, national et international s'appliquent déjà ou sont pertinentes dans le cadre de la mise au point, du déploiement et de l'utilisation de systèmes d'IA. Les sources de droit pertinentes comprennent, sans s'y limiter, le droit primaire de l'Union (les traités de l'Union européenne et sa charte des droits fondamentaux), le droit dérivé de l'Union (comme le règlement général sur la protection des données, les directives antidiscrimination, la directive «Machines», la directive sur la responsabilité du fait des produits, le règlement sur la libre circulation des données à caractère non personnel, les directives relatives au droit des consommateurs et à la sécurité et la santé au travail), mais également les traités en matière de droits de l'homme des Nations unies et les conventions du Conseil de l'Europe (telles que la Convention européenne des droits de l'homme) et de nombreuses autres législations des États membres de l'Union. Outre les règles applicables au niveau transversal, il existe différentes règles propres à un domaine donné qui s'appliquent à des applications d'IA particulières (comme le règlement relatif aux dispositifs médicaux dans le secteur des soins de santé).

(23) La législation prévoit des obligations tant positives que négatives; autrement dit, il convient de ne pas l'interpréter uniquement en lien avec ce qui ne *peut pas* être fait, mais aussi en lien avec ce qui *devrait* être fait. La législation ne se limite pas à interdire certaines actions mais en rend également d'autres possibles. À cet égard, il convient de noter que la charte de l'Union contient des articles relatifs à la «liberté d'entreprise» et à la «liberté des arts et des sciences», ainsi que des articles portant sur des domaines que nous connaissons mieux lorsqu'il s'agit de veiller à la fiabilité de l'IA, tels que la protection des données et la non-discrimination.

(24) Les lignes directrices ne traitent pas explicitement de la première caractéristique d'une IA digne de confiance (IA licite), mais visent plutôt à proposer des orientations pour encourager et garantir les deuxième et troisième caractéristiques (une IA éthique et robuste). Si ces deux dernières caractéristiques sont dans une certaine mesure déjà reflétées dans la législation existante, leur pleine réalisation pourrait aller au-delà des obligations juridiques existantes.

(25) Aucune partie du présent document ne peut s'entendre ou être interprétée comme fournissant des conseils ou orientations juridiques sur la manière de se mettre en conformité avec les normes et exigences juridiques existantes applicables. Aucun élément du présent document ne peut créer des droits ou imposer des obligations juridiques vis-à-vis de tiers. Nous rappelons toutefois que toute personne physique ou morale se doit de respecter la législation – qu'elle soit applicable aujourd'hui ou adoptée dans le futur en fonction de l'évolution de l'IA. Les présentes lignes directrices partent du principe que **l'ensemble des droits et obligations juridiques applicables aux processus et activités faisant partie de la mise au point, du déploiement et de l'utilisation de l'IA conservent un caractère obligatoire et doivent être dûment respectés.**

IA éthique

(26) Le respect du droit n'est qu'une des trois caractéristiques pour parvenir à la mise en œuvre d'une IA digne de confiance. La législation ne suit pas toujours le rythme des évolutions technologiques, ne correspond parfois pas à des normes éthiques ou peut simplement s'avérer inadaptée face à certaines questions. Pour être dignes de confiance, les systèmes d'IA devraient donc également être éthiques, en veillant à l'alignement sur les normes éthiques.

IA robuste

(27) Même lorsqu'une finalité éthique est garantie, les individus et la société doivent également être convaincus que les systèmes d'IA ne causeront pas de préjudice involontaire. Ces systèmes devraient être mis en œuvre de manière sûre, sécurisée et fiable, et il importe de prévoir des garanties pour éviter les incidences négatives involontaires. Il est par conséquent important de veiller à la robustesse des systèmes d'IA. Cet élément est nécessaire tant sur le plan technique (veiller à la robustesse technique du système selon les besoins dans un contexte donné, tel que le domaine d'application ou la phase du cycle de vie), que sur le plan social (en tenant

dûment compte du contexte et de l'environnement dans lesquels le système fonctionne). L'éthique et la robustesse de l'IA sont donc étroitement liées et se complètent mutuellement. Les principes mis en avant au chapitre I, ainsi que les exigences qui en découlent au chapitre II, portent sur ces deux caractéristiques.

Le cadre

(28) Les orientations du présent document se présentent sous la forme de trois niveaux d'abstraction, du plus abstrait, au chapitre I, au plus concret, au chapitre III:

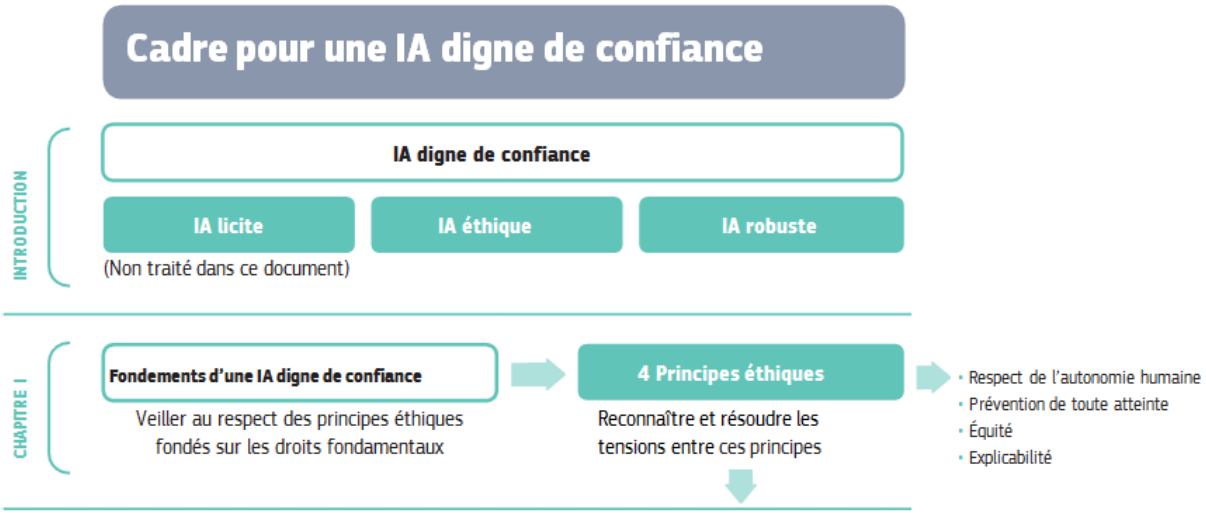
I) Fondements d'une IA digne de confiance. Le chapitre I établit les fondements d'une IA digne de confiance, en définissant son approche fondée sur le respect des droits fondamentaux[12]. Il recense et décrit les principes éthiques auxquels il convient d'adhérer afin de garantir une IA éthique et robuste.

II) Parvenir à une IA digne de confiance. Le chapitre II traduit ces principes éthiques en sept exigences que les systèmes d'IA devraient mettre en œuvre et respecter tout au long de leur cycle de vie. En outre, il propose des méthodes tant techniques que non techniques pouvant être appliquées aux fins de leur mise en œuvre.

III) Évaluer une IA digne de confiance. Les professionnels de l'IA attendent des orientations concrètes. Le chapitre III établit par conséquent une liste d'évaluation préliminaire et non exhaustive pour une IA digne de confiance afin de concrétiser les exigences du chapitre II. Cette évaluation devrait être adaptée à l'application spécifique du système.

(29) La dernière section du présent document expose des possibilités bénéfiques et des préoccupations importantes suscitées par les systèmes d'IA dont il convient de tenir compte et sur lesquelles nous souhaitons encourager de nouvelles discussions.

(30) La structure des présentes lignes directrices est illustrée à la *figure 1* ci-dessous.



[12] Les droits fondamentaux sont le fondement du droit tant international que de l'Union en matière de droits de l'homme et sous-tendent les droits opposables garantis par les traités de l'Union et par la charte des droits fondamentaux de l'Union européenne. Les droits fondamentaux étant juridiquement contraignants, leur respect relève donc de la première caractéristique d'une IA digne de confiance, à savoir une «IA licite». Les droits fondamentaux peuvent toutefois être interprétés comme reflétant aussi des droits moraux spéciaux reconnus à l'ensemble des individus en vertu de leur humanité, que ces droits soient ou non juridiquement contraignants. En ce sens, ils relèvent également de la deuxième caractéristique d'une IA digne de confiance, à savoir une «IA éthique».

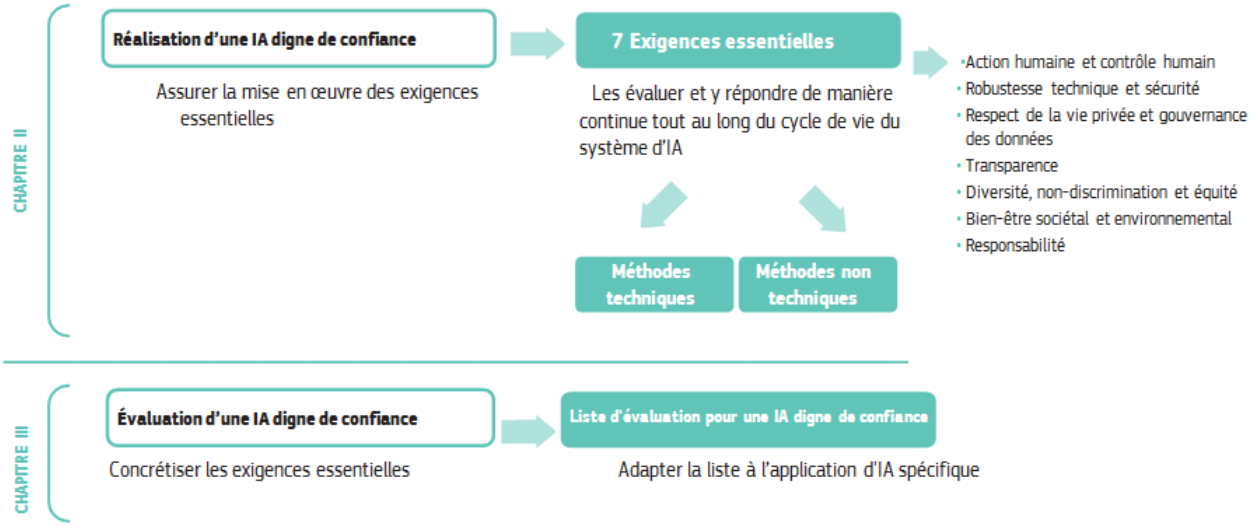


Figure 1: les lignes directrices en tant que cadre pour une IA digne de confiance

I. Chapitre I: Fondements d'une IA digne de confiance

(31) Ce chapitre établit les fondements d'une IA digne de confiance, reposant sur les droits fondamentaux et reflétée par quatre principes éthiques auxquels il convient d'adhérer afin de garantir une IA éthique et robuste. Ce chapitre s'appuie fortement sur le domaine de l'éthique.

(32) L'éthique en matière d'IA est un sous-domaine de l'éthique appliquée qui est axé sur les questions d'ordre éthique soulevées par la mise au point, le déploiement et l'utilisation de l'IA. Sa préoccupation centrale consiste à déterminer la manière dont l'IA peut soulever des préoccupations relatives au bien-être des individus ou y apporter des solutions, que ce soit du point de vue de la qualité de vie ou de l'autonomie humaine et de la liberté nécessaire pour une société démocratique.

(33) Une réflexion éthique sur la technologie de l'IA peut servir plusieurs objectifs. Premièrement, elle peut stimuler la réflexion sur la nécessité de protéger les individus et les groupes au niveau le plus élémentaire. Deuxièmement, elle peut stimuler de nouveaux genres d'innovation dont l'objectif est de promouvoir des valeurs éthiques, telles que celles contribuant à la réalisation des objectifs de développement durable des Nations unies[13], qui sont fermement ancrés dans le futur programme de l'Union européenne à l'horizon 2030[14]. Si le présent document porte principalement sur le premier objectif mentionné, il ne faut pas sous-estimer l'importance que pourrait revêtir l'éthique dans le cadre du deuxième objectif. Une IA digne de confiance peut renforcer la prospérité des individus et le bien-être collectif en générant de la prospérité, en créant de la valeur et en maximisant les richesses. Elle peut contribuer à la réalisation d'une société juste, en contribuant à l'amélioration de la santé et du bien-être des citoyens d'une manière qui renforce l'égalité dans la répartition des possibilités économiques, sociales et politiques.

(34) Il est par conséquent impératif que nous comprenions comment soutenir au mieux la mise au point, le déploiement et l'utilisation de l'IA pour faire en sorte que chacun puisse s'épanouir dans un monde fondé sur l'IA, et pour préparer un avenir meilleur tout en préservant la compétitivité au niveau mondial. Comme toute technologie puissante, l'utilisation de systèmes d'IA au sein de notre société soulève plusieurs problèmes éthiques, par exemple en ce qui concerne leur incidence sur les individus et la société, les capacités de prise de décision et la sécurité. Si nous prévoyons de nous faire assister par des systèmes d'IA ou de leur déléguer de plus en plus de décisions, nous devons veiller à ce que l'incidence de ces systèmes sur la vie des personnes soit équitable, à ce que ces systèmes soient conformes aux valeurs inaliénables et capables d'agir en ce sens, ainsi qu'à l'existence de processus adaptés en matière de responsabilisation pour y veiller.

(35) L'Europe doit définir la vision normative qu'elle souhaite mettre en œuvre pour un avenir marqué par l'omniprésence de l'IA et, par conséquent, comprendre quelle notion de l'IA devrait être étudiée, mise au point, déployée et utilisée en Europe pour réaliser cette vision. Avec ce document, nous souhaitons contribuer à cet effort en introduisant la notion d'IA digne de confiance qui est, selon nous, la manière adaptée de bâtir un avenir avec l'IA. Un avenir dans lequel la démocratie, l'état de droit et les droits fondamentaux sous-tendent les systèmes d'IA et dans lequel ces systèmes améliorent et défendent de manière continue la culture démocratique permettra également de mettre en place un environnement dans lequel l'innovation et la compétitivité responsable peuvent se développer.

(36) Un code de déontologie spécifique à un domaine donné – quel que soit le niveau de cohérence, d'élaboration et de détail de ses futures versions – ne pourra jamais se substituer à un raisonnement éthique en tant que tel, qui doit en toutes circonstances rester sensible aux éléments de contexte, qui ne peuvent jamais être rendus dans des lignes directrices générales. En plus d'élaborer un ensemble de règles, il convient de mettre sur pied et de conserver une culture et un état d'esprit éthiques dans le débat public, l'éducation et l'apprentissage pratique pour garantir une IA digne de confiance.

---

13 https://ec.europa.eu/commission/publications/reflection-paper-towards-sustainable-europe-2030_fr.

14 https://sustainabledevelopment.un.org/?menu=1300.

1. Les droits fondamentaux en tant que droits moraux et légaux

(37) Nous croyons en une approche de l'éthique en matière d'IA qui est fondée sur les droits fondamentaux consacrés par les traités de l'Union,[15] la charte des droits fondamentaux de l'Union européenne (charte de l'Union) et le droit international en matière de droits de l'homme.[16]Le respect des droits fondamentaux, dans le cadre de la démocratie et de l'état de droit, est le fondement le plus prometteur pour recenser les principes et les valeurs éthiques abstraits pouvant être concrétisés dans le contexte de l'IA.

(38) Les traités de l'Union et la charte de l'Union prescrivent un ensemble de droits fondamentaux que les États membres et les institutions de l'UE sont juridiquement tenus de respecter dans le cadre de la mise en œuvre du droit de l'Union. Ces droits sont décrits dans la charte de l'Union en référence à la dignité, aux libertés, à l'égalité et la solidarité, aux droits des citoyens et à la justice. Ces droits ont pour base commune un ancrage dans le respect de la dignité humaine, reflété par ce que nous décrivons comme une «approche centrée sur l'humain», dans laquelle l'être humain jouit d'un statut moral unique et inaliénable de primauté dans les domaines civil, politique, économique et social.[17]

(39) Alors que les droits établis dans la charte de l'Union sont juridiquement contraignants,[18] il est important de reconnaître que les droits fondamentaux n'assurent pas dans tous les cas une protection juridique complète. En ce qui concerne par exemple la charte de l'Union, il est important de souligner que son champ d'application se limite aux domaines couverts par le droit de l'Union. Le droit international en matière de droits de l'homme et notamment la Convention européenne des droits de l'homme sont juridiquement contraignants pour les États membres de l'UE, y compris dans les domaines qui sortent du champ d'application du droit de l'Union. Dans le même temps, il convient de souligner que des droits fondamentaux sont également conférés aux individus et (dans une certaine mesure) aux groupes en vertu de leur statut moral en tant qu'êtres humains, indépendamment de leur force juridique. Interprétés comme droits opposables, les droits fondamentaux relèvent par conséquent de la première caractéristique d'une IA digne de confiance (IA licite), qui garantit la conformité avec le droit. Interprétés comme les droits de chacun, ancrés dans le statut moral inhérent aux êtres humains, ils sous-tendent également la deuxième caractéristique d'une IA digne de confiance (IA éthique), qui porte sur des normes éthiques qui, sans être nécessairement contraignantes sur le plan juridique, sont pourtant essentielles pour parvenir à une IA digne de confiance. Étant donné que le présent document n'a pas vocation à fournir des orientations relatives à la première caractéristique, aux fins des présentes orientations non contraignantes, les références aux droits fondamentaux reflètent la deuxième caractéristique.

2. Des droits fondamentaux aux principes éthiques

2.1 Les droits fondamentaux comme base d'une IA digne de confiance

(40) Parmi l'éventail complet de droits indivisibles énoncés dans le droit international en matière de droits de l'homme, les traités de l'Union et la charte de l'Union, les familles de droits fondamentaux mentionnées ci-après sont particulièrement adaptées à une application aux systèmes d'IA. Une part importante de ces droits sont, dans des circonstances définies, opposables au sein de l'UE, ce qui rend juridiquement obligatoire la

[15] L'UE est fondée sur l'engagement constitutionnel de protéger les droits fondamentaux et indivisibles des êtres humains, de veiller au respect de l'état de droit, d'encourager la liberté démocratique et de promouvoir le bien commun. Ces droits sont reflétés aux articles 2 et 3 du traité sur l'Union européenne, ainsi que dans la charte des droits fondamentaux de l'Union européenne.

[16] D'autres instruments juridiques reflètent et précisent ces engagements, comme la charte sociale européenne du Conseil de l'Europe ou des législations spécifiques telles que le règlement général sur la protection des données de l'Union.

[17] Il convient de noter qu'un engagement envers une IA centrée sur l'humain et son ancrage dans les droits fondamentaux, plutôt que de supposer une valeur indûment individualiste de l'humain, nécessite des fondements sociétaux et constitutionnels collectifs dans lesquels la liberté individuelle et le respect de la dignité humaine sont à la fois possibles et pertinents.

[18] Conformément à l'article 51 de la charte, celle-ci s'applique aux institutions et aux États membres de l'Union lorsqu'ils mettent en œuvre le droit de l'Union.

conformité avec leurs exigences. Toutefois, même lorsque la conformité avec les droits fondamentaux opposables a été atteinte, une réflexion éthique peut nous aider à comprendre de quelle manière la mise au point, le déploiement et l'utilisation de l'IA peuvent mettre en jeu les droits fondamentaux et leurs valeurs sous-jacentes, et peuvent contribuer à des orientations plus précises lorsqu'il s'agit de déterminer ce que nous *devrions* faire plutôt que ce que nous *pouvons* faire (actuellement) à l'aide de la technologie.

(41) ***Respect de la dignité humaine***. La dignité humaine comprend l'idée que chaque être humain possède une «valeur intrinsèque», qui ne devrait jamais être diminuée, compromise ou réprimée par autrui – ni par de nouvelles technologies telles que des systèmes d'IA.[19] Dans le contexte de l'IA, le respect de la dignité humaine signifie que chaque personne est traitée avec respect du fait de son statut de *sujet* moral, plutôt que comme simple *objet* que l'on trie, classe, marque, régente, conditionne ou manipule. Les systèmes d'IA devraient donc être mis au point de manière à respecter, protéger et servir l'intégrité physique et mentale des êtres humains, leur sentiment d'identité personnel et culturel et la satisfaction de leurs besoins essentiels.[20]

(42) ***Liberté des individus.*** Les êtres humains devraient rester libres de faire leurs propres choix de vie. Cela suppose l'absence d'intrusion du pouvoir, mais requiert également l'intervention des pouvoirs publics et des organisations non gouvernementales pour faire en sorte que les individus exposés au risque d'exclusion jouissent d'une égalité d'accès aux avantages et aux possibilités que présente l'IA. Dans un contexte d'IA, la liberté des individus requiert l'atténuation des contraintes illégitimes, des menaces à l'encontre de l'autonomie et de la santé mentales, de la surveillance injustifiée, de la tromperie et de la manipulation injuste, que ces atteintes soient directes ou indirectes. En fait, la liberté des individus signifie un engagement visant à permettre aux individus d'exercer un contrôle accru sur leurs vies, y compris (entre autres droits) la protection de la liberté d'entreprise, la liberté des arts et des sciences, la liberté d'expression, le droit à la vie privée et à la confidentialité, et la liberté de réunion et d'association.

(43) ***Respect de la démocratie, de la justice et de l'état de droit.*** Dans les démocraties constitutionnelles, tout pouvoir gouvernemental doit être légalement autorisé et limité par la loi. Les systèmes d'IA devraient servir à conserver et à encourager les processus démocratiques et le respect de la pluralité des valeurs et des choix de vie des individus. Les systèmes d'IA ne doivent pas compromettre les processus démocratiques, la délibération humaine ou les systèmes de vote démocratiques. Les systèmes d'IA doivent également intégrer l'engagement de veiller à ce qu'ils ne soient pas mis en œuvre d'une manière qui compromette les engagements fondamentaux sur lesquels se fondent l'état de droit, les législations et règlements contraignants, et de garantir le droit à une procédure régulière et à l'égalité en droit.

(44) ***Égalité, non-discrimination et solidarité*** – *y compris le droit des personnes exposées au risque d'exclusion.* Il convient d'assurer un respect égal de la valeur morale et de la dignité de tous les êtres humains. Cela va au-delà de la non-discrimination, qui tolère le fait d'établir des distinctions entre des situations différentes sur la base de justifications objectives. Dans un contexte d'IA, l'égalité implique que le fonctionnement du système ne peut pas produire de résultats fondés sur des biais injustes (par exemple, les données utilisées pour entraîner les systèmes d'IA devraient être aussi inclusives que possible et représenter différents groupes de population), ce qui requiert également un respect approprié des personnes et des groupes potentiellement vulnérables[21], tels que les travailleurs, les femmes, les personnes handicapées, les minorités ethniques, les enfants, les consommateurs ou d'autres catégories de personnes exposées au risque d'exclusion.

(45) ***Droits des citoyens.*** Les citoyens bénéficient d'un large éventail de droits, dont le droit de vote, le droit à une bonne administration ou à l'accès aux documents publics, et le droit d'adresser des pétitions à l'administration. Grâce aux systèmes d'IA, les pouvoirs publics seront en mesure de fournir à la société des

[19] C. McCrudden, Human Dignity and Judicial Interpretation of Human Rights, *EJIL*, 19(4), 2008.

[20] Pour comprendre la «dignité humaine» en ce sens, voir E. Hilgendorf, Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity, dans: D. Grimm, A. Kemmerer, C. Möllers (eds.), *Human Dignity in Context. Explorations of a Contested Concept*, 2018, pp. 325 et suiv.

[21] Pour une description de ce terme tel qu'il est employé dans le présent document, voir le glossaire.

biens et des services publics à une échelle et avec une efficience supérieures. Dans le même temps, les applications d'IA pourraient aussi avoir une incidence négative sur les droits des citoyens; il convient par conséquent de protéger ces droits. L'emploi du terme «droits des citoyens» dans le présent document ne signifie nullement que nous nions ou négligeons les droits des ressortissants de pays tiers et des personnes en situation irrégulière (ou illégale) sur le territoire de l'UE, qui jouissent également de droits au titre du droit international et, par conséquent, dans le domaine de l'IA.

2.2 Principes éthiques dans le contexte des systèmes d'IA[22]

(46) De nombreuses organisations publiques, privées et civiles se sont inspirées des droits fondamentaux pour élaborer des cadres éthiques pour les systèmes d'IA.[23] Dans l'UE, le Groupe européen d'éthique des sciences et des nouvelles technologies («GEE») a proposé un ensemble de neuf principes fondamentaux, reposant sur les valeurs fondamentales énoncées dans les traités de l'Union et dans la charte des droits fondamentaux de l'Union européenne.[24] Nous continuons à nous appuyer sur ces travaux, en reconnaissant la plupart des principes avancés jusqu'à présent par différents groupes, tout en précisant à quelles fins l'ensemble de ces principes cherchent à répondre et à apporter un soutien. Ces principes éthiques peuvent inspirer de nouveaux instruments réglementaires spécifiques, contribuer à l'interprétation des droits fondamentaux au fur et à mesure qu'évolue notre environnement sociotechnique et orienter les motifs justifiant la mise au point, l'utilisation et la mise en œuvre de systèmes d'IA – en s'adaptant de manière dynamique aux évolutions de la société elle-même.

(47) Les systèmes d'IA doivent améliorer le bien-être individuel et collectif. Cette section présente **quatre principes éthiques**, ancrés dans les droits fondamentaux, auxquels il convient d'adhérer pour faire en sorte que les systèmes d'IA soient mis au point, déployés et utilisés d'une manière digne de confiance. Ils sont présentés comme des **impératifs éthiques**, si bien que les professionnels de l'IA devraient en toutes circonstances s'efforcer d'y adhérer. Sans imposer de hiérarchie, nous présentons les principes ci-dessous de manière à refléter l'ordre d'apparition, dans la charte de l'Union, des droits fondamentaux sur lesquels ils se fondent.[25]

(48) Il s'agit des principes suivants:

(i) respect de l'autonomie humaine
(ii) prévention de toute atteinte
(iii) équité
(iv) explicabilité

(49) La plupart de ces principes sont dans une large mesure déjà reflétés dans les exigences juridiques contraignantes dont la mise en œuvre est obligatoire et relèvent donc également du champ d'application de l'«IA licite», soit la première caractéristique d'une IA digne de confiance.[26] Pourtant, comme indiqué plus haut, même si de nombreuses obligations juridiques reflètent des principes éthiques, l'adhésion à des principes

[22] Ces principes s'appliquent également à la mise au point, au déploiement et à l'utilisation d'autres technologies, et ne sont par conséquent pas spécifiques aux systèmes d'IA. Nous nous sommes efforcés ci-dessous d'établir leur pertinence dans un contexte spécifiquement lié à l'IA.

[23] Le recours aux droits fondamentaux contribue également à limiter l'insécurité réglementaire, car elle peut s'appuyer sur des décennies de pratique en matière de protection des droits fondamentaux dans l'UE, ce qui apporte de la clarté, de la lisibilité et de la prévisibilité.

[24] Plus récemment, le groupe de travail de AI4People a examiné les principes susmentionnés du GEE ainsi que 36 autres principes éthiques énoncés à ce jour et les a ramenés à quatre principes généraux. L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena (2018), «AI4People —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations», Minds and Machines 28(4): p. 689-707.

[25] Le respect de l'autonomie humaine est fortement associé au droit à la dignité humaine et à la liberté (reflété aux articles 1 et 6 de la charte). La prévention de toute atteinte est fortement liée à la protection de l'intégrité physique ou mentale (reflétée à l'article 3). L'équité est étroitement liée aux droits à la non-discrimination, à la solidarité et à la justice (reflétés aux articles 21 et suivants). L'explicabilité et la responsabilité sont étroitement liées aux droits relatifs à la justice (tels que reflétés à l'article 47).

[26] On pense par exemple au RGPD ou aux règlements de l'Union relatifs à la protection des consommateurs.

éthiques dépasse le respect formel de la législation existante.[27]

- Le principe du respect de l'autonomie humaine

(50) Les droits fondamentaux sur lesquels l'UE est fondée ont vocation à garantir le respect de la liberté et de l'autonomie des êtres humains. Les êtres humains qui interagissent avec des systèmes d'IA doivent être en mesure de conserver leur autodétermination totale et effective et de prendre part au processus démocratique. En l'absence de justification, les systèmes d'IA ne devraient pas subordonner, contraindre, tromper, manipuler, conditionner ni régenter des êtres humains. Au contraire, les systèmes d'IA devraient être conçus afin d'augmenter, de compléter et de favoriser les compétences cognitives, sociales et culturelles. La répartition des tâches entre êtres humains et systèmes d'IA devrait suivre des principes de conception centrés sur l'humain et donner à l'être humain une possibilité réelle de poser des choix. En d'autres termes, il convient de veiller à la supervision[28] et au contrôle humains sur les processus de travail des systèmes d'IA. Les systèmes d'IA pourraient également modifier fondamentalement la sphère du travail. Ces systèmes devraient aider les êtres humains dans l'environnement de travail, et avoir pour objectif de créer des emplois qui aient du sens.

- Le principe de la prévention de toute atteinte

(51) Les systèmes d'IA ne devraient ni porter atteinte, ni aggraver toute atteinte portée[29], ni nuire aux êtres humains d'une quelconque autre manière.[30] Cela englobe la protection de la dignité humaine ainsi que de l'intégrité mentale et physique. Les systèmes d'IA et les environnements dans lesquels ils évoluent doivent être sûrs et sécurisés. Ils doivent être robustes sur le plan technique et il convient de veiller à ce qu'ils ne soient pas exposés à des utilisations malveillantes. Les personnes vulnérables devraient faire l'objet d'une attention accrue et être prises en compte dans la mise au point et le déploiement des systèmes d'IA. Il convient également d'accorder une attention particulière aux situations dans lesquelles les systèmes d'IA peuvent entraîner ou aggraver des incidences négatives du fait d'asymétries de pouvoir ou d'information, par exemple entre les employeurs et les travailleurs, entre les entreprises et les consommateurs ou entre les pouvoirs publics et les citoyens. La prévention de toute atteinte implique également la prise en compte de l'environnement naturel et de tous les êtres vivants.

- Le principe de l'équité

(52) La mise au point, le déploiement et l'utilisation de systèmes d'IA doivent être équitables. Si nous reconnaissons que l'équité peut s'interpréter de multiples manières, nous considérons que l'équité se caractérise à la fois par un volet matériel et un volet procédural. Le volet matériel suppose l'engagement de veiller à une répartition égale et juste des bénéfices et des coûts, et de veiller à ce que les individus et les groupes ne fassent pas l'objet de biais injustes, de discrimination et de stigmatisation. Si les biais injustes peuvent être évités, les systèmes d'IA pourraient même améliorer le caractère équitable de la société. Il convient également d'encourager l'égalité des chances en ce qui concerne l'accès à l'éducation, aux biens, aux services et à la technologie. En outre, l'utilisation de systèmes d'IA ne devrait jamais avoir pour conséquence de tromper les utilisateurs (finaux) ou de limiter leur liberté de choix. L'équité implique en outre que les professionnels de l'IA devraient respecter le principe de proportionnalité entre la fin et les moyens, et examiner de manière attentive la manière de trouver un équilibre entre des intérêts et des objectifs en

---

27 Pour d'autres références sur le sujet, voir par exemple L. Floridi, Soft Ethics and the Governance of the Digital, *Philosophy & Technology,* March 2018, Volume 31, Issue 1, pp 1–8.

28 Le concept du contrôle humain est approfondi au point 65 ci-dessous.

29 Une atteinte portée peut être individuelle ou collective, et peut comprendre une atteinte immatérielle aux environnements sociaux, culturels et politiques.

30 Les atteintes au mode de vie des individus et des groupes sociaux peuvent être qualifiées d'atteintes culturelles et doivent être évitées.

concurrence.[31] Le volet procédural de l'équité suppose la capacité de contester les décisions prises par des systèmes d'IA et par les êtres humains qui les utilisent, ainsi que celle d'introduire un recours efficace à l'encontre de ces décisions[32]. Pour ce faire, l'entité responsable de la décision doit pouvoir être identifiée, et le processus de prise de décisions devrait pouvoir être expliqué.

- Le principe de l'explicabilité

(53) L'explicabilité est essentielle pour renforcer et conserver la confiance des utilisateurs envers les systèmes d'IA. Cela signifie que les processus doivent être transparents, que les capacités et la finalité des systèmes d'IA doivent être communiquées ouvertement, et que les décisions – dans la mesure du possible – doivent pouvoir être expliquées aux personnes directement et indirectement concernées. Sans ces informations, une décision ne peut être dûment contestée. Il n'est pas toujours possible d'expliquer pour quelle raison un modèle a généré un résultat ou une décision en particulier (et quelle combinaison de facteurs d'entrée y a contribué). On parle d'algorithmes à effet «boîte noire». Ceux-ci doivent faire l'objet d'une attention particulière. Dans de telles circonstances, d'autres mesures d'explicabilité (par exemple la traçabilité, l'auditabilité et la communication transparente concernant les capacités du système) pourraient être requises, pour autant que le système dans son ensemble respecte les droits fondamentaux. La mesure dans laquelle l'explicabilité est nécessaire dépend fortement du contexte et de la gravité des conséquences si ce résultat est erroné ou imprécis d'une autre manière.[33]

2.3 Tensions entre ces principes

(54) Des tensions pourraient survenir entre les principes susmentionnés, pour lesquelles il n'existe pas de solution unique. En vertu de l'engagement fondamental de l'UE envers l'engagement démocratique, le droit à une procédure régulière et la participation politique ouverte, des méthodes de délibération responsable devraient être établies pour faire face à ces tensions. Par exemple, dans divers domaines d'application, *le principe de la prévention de toute atteinte* et *le principe de l'autonomie humaine* peuvent entrer en conflit. Ainsi, l'utilisation de systèmes d'IA aux fins d'une «police prédictive» pourrait contribuer à réduire la criminalité, mais d'une manière impliquant des activités de surveillance qui portent atteinte à la liberté individuelle et à la vie privée. En outre, la somme des avantages liés aux systèmes d'IA doit être sensiblement supérieure aux risques individuels prévisibles. Si ces principes fournissent clairement des orientations destinées à trouver des solutions, ils n'en demeurent pas moins des prescriptions éthiques abstraites. On ne peut attendre des professionnels de l'IA qu'ils trouvent la solution adaptée sur la base des principes ci-dessus. Il leur faut toutefois aborder les dilemmes et arbitrages éthiques selon une réflexion raisonnée et fondée sur des éléments probants, plutôt que sur la base de l'intuition ou d'un jugement aléatoire. Il pourrait toutefois exister des situations dans lesquelles aucun arbitrage acceptable du point de vue éthique ne peut être déterminé. Certains droits fondamentaux et principes connexes sont absolus et ne peuvent dépendre d'un exercice de mise en balance (par exemple, la dignité humaine).

Orientations essentielles dérivées du chapitre I:

[31] Cette exigence est liée au principe de la proportionnalité (reflété par la maxime selon laquelle «on ne tue pas une mouche avec un bazooka»). Les mesures prises pour parvenir à une fin (par exemple, l'extraction de données en vue d'optimiser l'IA) devraient être limitées au strict nécessaire. Cela implique également que lorsque plusieurs mesures sont en concurrence pour la réalisation d'un même but, la préférence devrait être accordée à celle qui est la moins défavorable aux droits fondamentaux et aux normes éthiques (par exemple, les développeurs d'IA devraient toujours accorder la préférence à des données du secteur public par rapport aux données à caractère personnel). Il convient également de faire référence à la proportionnalité entre l'utilisateur et le prestataire du déploiement, en tenant compte des droits des entreprises (y compris de propriété intellectuelle et de confidentialité), d'une part, et des droits de l'utilisateur, d'autre part.

[32] Notamment en invoquant leur droit d'association et d'adhérer à un syndicat dans un environnement de travail, comme le prévoit l'article 12 de la charte des droits fondamentaux de l'Union européenne.

[33] Par exemple, les préoccupations éthiques résultant de recommandations d'achat imprécises générées par un système d'IA ne pourraient être que limitées, contrairement à celles résultant de systèmes d'IA évaluant si un individu reconnu coupable d'une infraction pénale devrait être mis en liberté conditionnelle.

- ✓ Mettre au point, déployer et utiliser des systèmes d'IA en respectant les principes éthiques suivants: *respect de l'autonomie humaine, prévention de toute atteinte, équité et explicabilité.* Reconnaître et résoudre les tensions potentielles entre ces principes.
- ✓ Accorder une attention particulière aux situations concernant des groupes plus vulnérables tels que les enfants, les personnes handicapées et d'autres groupes historiquement défavorisés, exposés au risque d'exclusion, et/ou aux situations caractérisées par des asymétries de pouvoir ou d'information, par exemple entre les employeurs et les travailleurs, ou entre les entreprises et les consommateurs.[34]
- ✓ Reconnaître et être conscient que certaines applications d'IA sont certes susceptibles d'apporter des avantages considérables aux individus et à la société, mais qu'elles peuvent également avoir des incidences négatives, y compris des incidences pouvant s'avérer difficiles à anticiper, reconnaître ou mesurer (par exemple, en matière de démocratie, d'état de droit et de justice distributive, ou sur l'esprit humain lui-même). Adopter des mesures appropriées pour atténuer ces risques le cas échéant, de manière proportionnée à l'ampleur du risque.

II. Chapitre II: Parvenir à une IA digne de confiance

(55) Ce chapitre fournit des orientations relatives à la mise en œuvre et à la réalisation d'une IA digne de confiance, au moyen d'une liste de sept exigences qui devraient être respectées, s'appuyant sur les principes énoncés au chapitre I. En outre, des méthodes tant techniques que non techniques actuellement disponibles sont présentées aux fins de l'application de ces exigences tout au long du cycle de vie du système d'IA.

1. Exigences d'une IA digne de confiance

(56) Pour parvenir à une IA digne de confiance, il faut que les principes énoncés au chapitre I soient traduits en exigences concrètes. Ces exigences s'appliquent aux différentes parties prenantes participant au cycle de vie des systèmes d'IA: développeurs, prestataires et utilisateurs finaux, ainsi que la société au sens large. Le terme «développeurs» désigne les personnes qui effectuent des recherches sur les systèmes d'IA, et qui conçoivent et/ou mettent au point ces systèmes. Le terme «prestataires» désigne les organismes publics ou privés qui utilisent des systèmes d'IA dans leurs processus opérationnels pour proposer des produits et services à des tiers. Les utilisateurs finaux sont les personnes qui interagissent directement ou indirectement avec le système d'IA. Enfin, la société au sens large englobe tous les autres acteurs qui sont directement ou indirectement concernés par les systèmes d'IA.

(57) Différentes catégories de parties prenantes ont différents rôles à jouer pour veiller au respect des exigences:

a. Les développeurs devraient mettre en œuvre et appliquer les exigences aux processus de conception et de mise au point;
b. Les prestataires devraient veiller à ce que les systèmes qu'ils utilisent et les produits et services qu'ils proposent respectent les exigences;
c. Les utilisateurs finaux et la société au sens large devraient être informés de ces exigences et être en mesure de demander qu'elles soient respectées.

(58) La liste des exigences ci-dessous n'est pas exhaustive.[35] Elle comprend des aspects systémiques, individuels et sociétaux:

1 Action humaine et contrôle humain

Comprend les droits fondamentaux, l'action humaine et le contrôle humain

[34] Voir articles 24 à 27 de la charte l'UE, portant sur les droits de l'enfant et des personnes âgées, l'intégration des personnes handicapées et le droit des travailleurs. Voir également l'article 38 portant sur la protection des consommateurs.

[35] Sans imposer de hiérarchie, nous présentons les principes ci-dessous de manière à refléter l'ordre d'apparition, dans la charte de l'Union, des principes et des droits auxquels ils se rapportent.

2 **Robustesse technique et sécurité**

Comprend la résilience aux attaques et la sécurité, les plans de secours et la sécurité générale, la précision, la fiabilité et la reproductibilité

3 **Respect de la vie privée et gouvernance des données**

Comprend le respect de la vie privée, la qualité et l'intégrité des données, et l'accès aux données

4 **Transparence**

Comprend la traçabilité, l'explicabilité et la communication

5 **Diversité, non-discrimination et équité**

Comprend l'absence de biais injustes, l'accessibilité et la conception universelle, et la participation des parties prenantes

6 **Bien-être sociétal et environnemental**

Comprend la durabilité et le respect de l'environnement, l'impact social, la société et la démocratie

7 **Responsabilité**

Comprend l'auditabilité, la réduction au minimum des incidences négatives et la communication à leur sujet, les arbitrages et les recours.

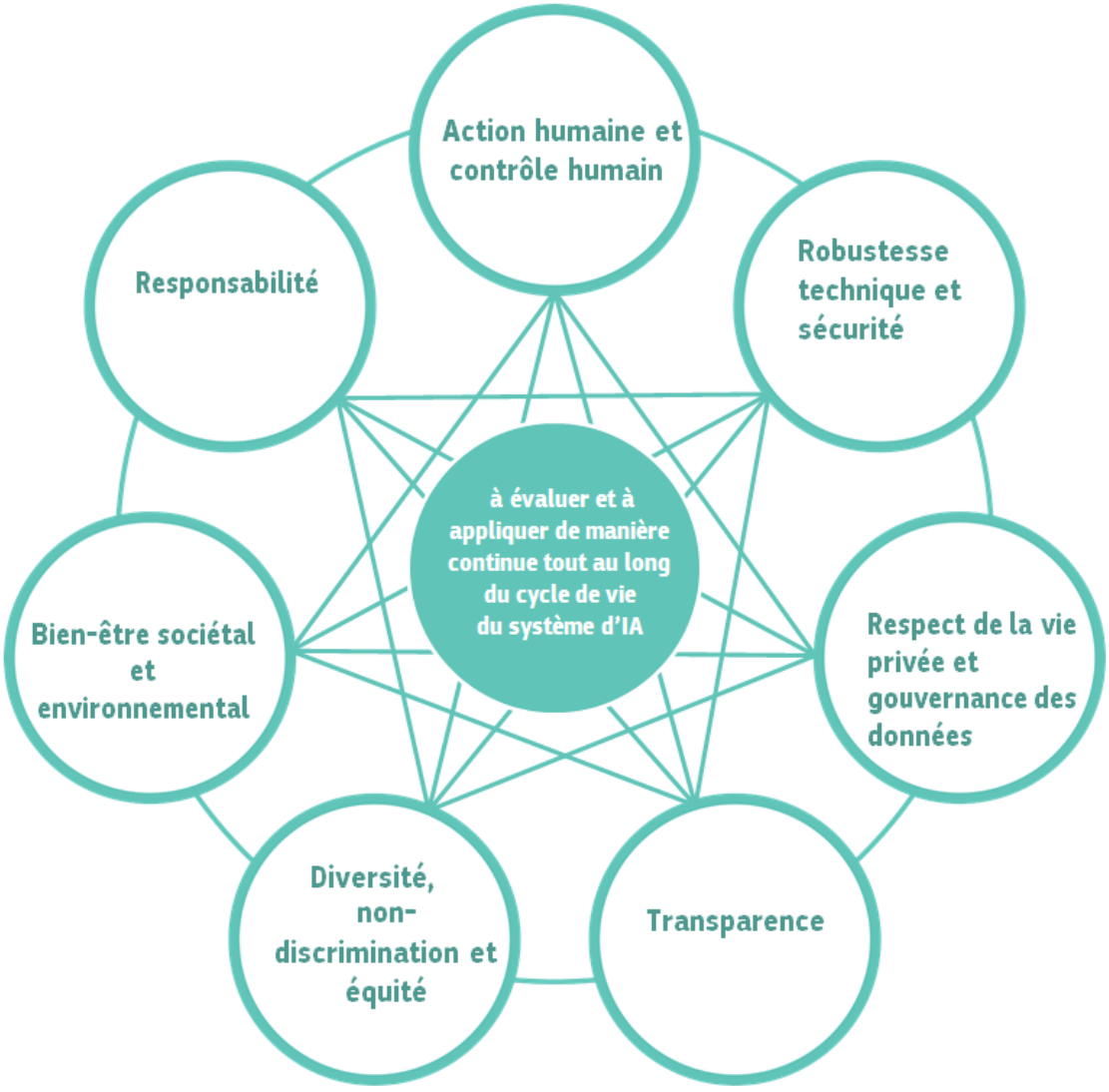


Figure 2: interrelation des sept exigences: elles revêtent toutes une importance égale, elles se soutiennent mutuellement et devraient être appliquées et évaluées tout au long du cycle de vie d'un système d'IA.

(59) Si toutes ces exigences revêtent une importance égale, le contexte et les tensions s'exerçant potentiellement entre elles devront être pris en compte lors de leur application à différents domaines et secteurs d'activité. La mise en œuvre de ces exigences devrait se faire tout au long du cycle de vie d'un système d'IA, et dépend de l'application spécifique. Si la plupart des exigences s'appliquent à l'ensemble des systèmes d'IA, une attention spécifique est accordée à celles qui ont des effets directs ou indirects sur les personnes. Par conséquent, pour certaines applications (par exemple, dans des contextes industriels), elles peuvent s'avérer moins pertinentes.

(60) Les exigences ci-dessus comprennent des éléments qui, dans certains cas, sont déjà reflétés dans la législation existante. Nous rappelons que – conformément à la première caractéristique d'une IA digne de confiance – les développeurs et prestataires de systèmes d'IA ont la responsabilité de faire en sorte qu'ils respectent leurs obligations juridiques, tant en ce qui concerne les règles applicables au niveau transversal que les règles spécifiques à un domaine donné.

(61) Dans les paragraphes qui suivent, chaque exigence fait l'objet d'un examen plus approfondi.

1. Action humaine et contrôle humain

(62) Les systèmes d'IA devraient soutenir l'autonomie et la prise de décisions humaines, conformément au principe du *respect de l'autonomie humaine*, en vertu duquel les systèmes d'IA devraient être à la fois les vecteurs d'une société démocratique, prospère et équitable en se mettant au service de l'utilisateur et favoriser les droits fondamentaux, ainsi que permettre un contrôle humain.

(63) ***Droits fondamentaux.*** À l'instar de nombreuses technologies, les systèmes d'IA peuvent autant favoriser qu'entraver les droits fondamentaux. Ils peuvent par exemple servir les particuliers en les aidant à suivre leurs données à caractère personnel ou en améliorant l'accès à l'éducation, en soutenant ainsi leur droit à l'éducation. Toutefois, étant donné la portée et la capacité des systèmes d'IA, ils peuvent également avoir une incidence négative sur les droits fondamentaux. Dans les situations où de tels risques existent, il convient d'entreprendre une analyse d'impact relative aux droits fondamentaux. Cette analyse devrait être menée préalablement à leur mise au point et comprendre une évaluation destinée à déterminer si ces risques peuvent être réduits ou justifiés comme nécessaires dans une société démocratique afin de respecter les droits et les libertés d'autrui. Il convient en outre de mettre sur pied des mécanismes permettant de recevoir des commentaires externes concernant les systèmes d'IA susceptibles de nuire aux droits fondamentaux.

(64) ***Action humaine.*** Les utilisateurs devraient être en mesure de prendre des décisions autonomes éclairées à l'égard des systèmes d'IA. Ils devraient recevoir les connaissances et les outils pour comprendre les systèmes d'IA et interagir avec eux dans une mesure satisfaisante et, autant que possible, être à même de procéder à une autoévaluation du système ou de le contester d'une manière appropriée. Les systèmes d'IA devraient aider les individus à prendre de meilleures décisions et à faire des choix plus éclairés en rapport avec leurs objectifs. Les systèmes d'IA peuvent parfois être déployés pour modeler et influencer le comportement humain à travers des mécanismes parfois difficiles à détecter, du fait qu'ils peuvent exploiter des processus subconscients, y compris différentes formes de manipulation déloyale, de tromperie, d'asservissement et de conditionnement, chacune étant susceptible de menacer l'autonomie individuelle. Le principe général d'autonomie des utilisateurs doit être au cœur des fonctionnalités du système. À cet égard, le droit des utilisateurs de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé lorsque cela produit sur eux des effets juridiques ou d'autres effets d'importance comparable [36] revêt un caractère essentiel.

(65) ***Contrôle humain***. Le contrôle humain contribue à éviter qu'un système d'IA ne mette en péril l'autonomie humaine ou ne provoque d'autres effets néfastes. Le contrôle peut être assuré en recourant à des mécanismes de gouvernance tels que les approches dites «human-in-the-loop» (l'humain intervient dans le processus),

[36] Il peut être fait référence à l'article 22 du RGPD qui consacre déjà ce droit.

«human-on-the-loop» (l'humain supervise le processus) ou «human-in-command» (l'humain reste aux commandes). L'approche «human-in-the-loop» (HITL) désigne la capacité d'intervention humaine dans chaque cycle de décision du système, ce qui, dans de nombreux cas, n'est ni possible ni souhaitable. L'approche «human-on-the-loop» (HOTL) désigne une capacité d'intervention humaine dans le cycle de conception du système et la surveillance du fonctionnement du système. L'approche «human-in-command» (HIC) désigne une capacité de contrôle de l'activité globale du système d'IA (y compris de ses incidences économiques, sociétales, juridiques et éthiques au sens large) et la faculté de décider quand et comment utiliser le système dans une situation donnée. Cette faculté peut comprendre la décision de ne pas utiliser un système d'IA dans une situation donnée, de définir des marges d'appréciation pour les interventions humaines lors de l'utilisation du système ou d'ignorer une décision prise par un système. Il convient en outre de veiller à ce que les autorités publiques soient en mesure d'exercer un contrôle conformément à leur mandat. Des mécanismes de contrôle peuvent être requis à des degrés divers pour soutenir d'autres mesures de sécurité et de contrôle, en fonction du domaine d'application du système d'IA et du risque potentiel. Toutes choses étant égales par ailleurs, moins un être humain peut exercer de contrôle sur un système d'IA, plus il faut approfondir les essais et renforcer la gouvernance.

2. Robustesse technique et sécurité

(66) Une caractéristique essentielle pour parvenir à une IA digne de confiance est la robustesse technique, qui est étroitement liée au *principe de la prévention de toute atteinte*. La robustesse technique passe par la mise au point de systèmes d'IA selon une approche de prévention des risques, et de telle manière que ces systèmes se comportent, de manière fiable, conformément aux attentes, tout en réduisant le plus possible les atteintes involontaires et inattendues, et en empêchant toute atteinte inacceptable. Cette exigence également s'appliquer aux modifications potentielles de l'environnement dans lequel ils sont exploités ou à la présence d'autres agents (humains et artificiels) pouvant avoir des interactions antagonistes avec le système. Il convient en outre de garantir l'intégrité physique et mentale des êtres humains.

(67) ***Résilience aux attaques et sécurité***. Les systèmes d'IA, à l'instar de tous les systèmes logiciels, devraient être protégés face aux vulnérabilités qui pourraient permettre à des adversaires de les exploiter (par exemple, piratage). Des attaques pourraient cibler les données (empoisonnement des données), le modèle (fuite de modèle) ou l'infrastructure sous-jacente, tant matérielle que logicielle. Lorsqu'un système d'IA fait l'objet d'une attaque, par exemple d'une attaque antagoniste, le comportement des données ainsi que du système peut être modifié, ce qui conduit le système à prendre des décisions différentes voire à s'arrêter. Les systèmes et les données peuvent également être corrompus en raison d'interventions malveillantes ou de l'exposition à des situations imprévues. Des procédures de sécurité insuffisantes peuvent également mener à des décisions erronées ou même entraîner des préjudices physiques. Pour que les systèmes d'IA soient considérés comme sûrs,[37] il convient de prendre en compte les applications involontaires potentielles de l'IA (par exemple, applications à double usage) et l'utilisation potentiellement abusive d'un système d'IA par des acteurs malveillants et de prendre des mesures pour les empêcher et les atténuer.[38]

(68) ***Plans de secours et sécurité générale***. Les systèmes d'IA devraient comporter des garanties permettant le déclenchement de plans de secours en cas de problèmes. Un système d'IA pourrait ainsi être amené à passer d'une procédure statistique à une procédure fondée sur des règles, ou à demander les instructions d'un

[37] Voir par exemple les considérations au point 2.7 du plan coordonné de l'Union européenne dans le domaine de l'intelligence artificielle.

[38] Pour assurer la sécurité des systèmes d'IA, il pourrait être indispensable de mettre en place un cercle vertueux en matière de recherche et de développement entre la compréhension des attaques, la mise au point de protections appropriées et l'amélioration des méthodes d'évaluation. Pour y parvenir, il convient de promouvoir une convergence entre la communauté de l'IA et la communauté de la sécurité. Il incombe en outre à l'ensemble des acteurs concernés de définir des normes communes de sûreté et de sécurité transfrontières et de mettre en place un environnement de confiance mutuelle, encourageant la collaboration internationale. Pour des mesures possibles, voir Malicious Use of AI (Avin S., Brundage M., et. al., 2018).

opérateur humain avant de poursuivre son action.[39] Il convient de s'assurer que le système fera ce qui est attendu de lui sans porter atteinte à des êtres vivants ou à l'environnement. Cela comprend la nécessité de réduire le plus possible les effets non désirés et les dysfonctionnements. En outre, des processus devraient être mis en place pour clarifier et évaluer les risques potentiels liés à l'utilisation de systèmes d'IA pour un éventail de domaines d'application. Le niveau des mesures de sécurité nécessaires dépend de l'ampleur du risque que présente un système d'IA, qui dépend en retour des capacités du système. Lorsqu'il apparaît prévisible que le processus de mise au point ou le système même présenteront des risques particulièrement élevés, il est essentiel de mettre au point et de tester de manière proactive des mesures de sécurité.

(69) ***Précision.*** La précision est fonction de la capacité d'un système d'IA à poser un jugement correct, par exemple en classant correctement des informations dans les bonnes catégories, ou de sa capacité à réaliser des prévisions, des recommandations ou des décisions correctes sur la base de données ou de modèles. Un processus de mise au point et d'évaluation explicite et bien formé peut, en plus d'apporter le soutien nécessaire, atténuer et corriger les risques imprévus découlant de prévisions inexactes. Lorsqu'il n'est pas possible d'éviter des prévisions inexactes occasionnelles, il est important que le système puisse indiquer le niveau de probabilité de ces erreurs. Un niveau élevé de précision est particulièrement essentiel dans les situations où le système d'IA a une incidence directe sur des vies humaines.

(70) ***Fiabilité et reproductibilité.*** Il est essentiel que les résultats des systèmes d'IA soient à la fois reproductibles et fiables. Un système d'IA fiable est un système qui fonctionne correctement avec toute une gamme de données d'entrée et dans un ensemble de situations. Ces caractéristiques sont nécessaires pour qu'un système d'IA puisse être soumis à un examen attentif et éviter tout préjudice involontaire. La reproductibilité est une indication de la mesure dans laquelle un système d'IA, dans le cadre d'essais répétés dans les mêmes conditions, produit un comportement similaire. Cela permet aux scientifiques et aux décideurs politiques de décrire avec précision ce que font les systèmes d'IA. Les fichiers de reproduction[40] peuvent faciliter le processus d'essai et de reproduction des comportements.

3. Respect de la vie privée et gouvernance des données

(71) Étroitement lié au *principe de la prévention de toute atteinte*, le respect de la vie privée est un droit fondamental particulièrement sensible aux incidences des systèmes d'IA. La prévention de toute atteinte au respect de la vie privée requiert également une gouvernance appropriée des données qui porte sur la qualité et l'intégrité des données utilisées, leur pertinence par rapport au domaine dans lequel les systèmes d'IA seront déployés, leurs protocoles d'accès et la capacité à traiter les données d'une manière qui protège la vie privée.

(72) ***Respect de la vie privée et protection des données***. Les systèmes d'IA doivent garantir le respect de la vie privée et la protection des données tout au long du cycle de vie d'un système.[41] Cela couvre les informations initialement fournies par l'utilisateur, ainsi que les informations générées au sujet de l'utilisateur au cours de ses interactions avec le système (par exemple, des résultats générés par le système d'IA pour des utilisateurs spécifiques, ou la manière dont les utilisateurs ont répondu à des recommandations spécifiques). La numérisation des comportements humains peut permettre aux systèmes d'IA de déduire non seulement les préférences d'une personne, mais aussi son orientation sexuelle, son âge, son sexe, ses convictions religieuses ou ses opinions politiques. Pour que les citoyens aient confiance dans le processus de collecte des données, ils doivent avoir la garantie que les données recueillies les concernant ne seront pas utilisées à leur encontre à des fins discriminatoires, de manière illicite ou injuste.

---

[39] Il convient également d'envisager des scénarios dans lesquels une intervention humaine ne serait pas immédiatement possible.

[40] Il s'agit de fichiers qui reproduiront chaque étape du processus de mise au point du système d'IA, du stade de la recherche et de la collecte initiale des données jusqu'au stade des résultats.

[41] Il peut être fait référence à la législation existante en matière de respect de la vie privée, telle que le RGPD ou le futur règlement «vie privée et communications électroniques».

(73) ***Qualité et intégrité des données***. La qualité des ensembles de données utilisés est essentielle au bon fonctionnement des systèmes d'IA. La collecte de données peut être entachée de biais d'ordre social, d'imprécisions, de fautes et d'erreurs. Il faut tenir compte de cet élément avant d'utiliser un ensemble de données pour entraîner un système d'IA. Par ailleurs, l'intégrité des données doit être assurée. Alimenter un système d'IA avec des données malveillantes peut modifier son comportement, notamment avec les systèmes d'autoapprentissage. Les processus et ensembles de données utilisés doivent être testés et documentés à chaque étape (planification, entraînement, essais et déploiement). Ce principe devrait s'appliquer également aux systèmes d'IA qui n'ont pas été développés en interne mais qui ont été acquis à l'extérieur.

(74) ***Accès aux données****.* Dans toute organisation traitant les données relatives à des personnes (qu'il s'agisse ou non d'utilisateurs du système), des protocoles de données régissant l'accès aux données devraient être mis en place. Ces protocoles devraient indiquer qui peut avoir accès aux données et dans quelles circonstances. Seul le personnel dûment qualifié ayant les compétences nécessaires et justifiant du besoin d'accéder à des données à caractère personnel devrait y être autorisé.

4. Transparence

(75) Cette exigence est étroitement liée au *principe de l'explicabilité* et comprend la transparence des éléments pertinents d'un système d'IA: les données, le système et les modèles économiques.

(76) ***Traçabilité****.* Les ensembles de données et les processus permettant au système d'IA de rendre une décision, y compris les processus de collecte et d'étiquetage de données, ainsi que les algorithmes utilisés, devraient être documentés selon les normes les plus strictes afin de permettre la traçabilité ainsi qu'une amélioration de la transparence. Ce principe s'applique également aux décisions rendues par le système d'IA. Cela permet de déterminer les raisons pour lesquelles une décision d'IA était erronée ce qui, en retour, pourrait contribuer à éviter de futures erreurs. La traçabilité facilite donc l'auditabilité et l'explicabilité.

(77) ***Explicabilité.*** L'explicabilité concerne la capacité d'expliquer à la fois les processus techniques d'un système d'IA et les décisions humaines qui s'y rapportent (par exemple, domaines d'application d'un système d'IA). L'explicabilité technique suppose que les décisions prises par un système d'IA puissent être comprises et retracées par des êtres humains. Par ailleurs, des arbitrages peuvent s'avérer nécessaires entre le renforcement de l'explicabilité d'un système (qui pourrait réduire sa précision) et l'amélioration de sa précision (au détriment de l'explicabilité). Dès qu'un système d'IA a une incidence importante sur la vie des personnes, il devrait être possible d'exiger une explication appropriée du processus de décision du système d'IA. Ces explications devraient être présentées en temps opportun et adaptées à l'expertise de la partie prenante concernée (par exemple, non-spécialiste, autorité de réglementation ou chercheur). Des explications devraient également être fournies sur la mesure dans laquelle un système d'IA influence et façonne le processus de prise de décisions organisationnel, les choix opérés dans la conception du système, et la justification de son déploiement (de manière à assurer la transparence du modèle économique).

(78) ***Communication.*** Les systèmes d'IA ne devraient pas se présenter comme des êtres humains auprès des utilisateurs; lorsqu'ils interagissent avec un système d'IA, les êtres humains ont le droit d'en être informés. Cet aspect implique que les systèmes d'IA doivent être identifiables en tant que tels. Qui plus est, la possibilité de s'opposer à cette interaction au profit d'une interaction humaine devrait être proposée le cas échéant afin de garantir le respect des droits fondamentaux. Outre cet aspect, il convient de communiquer aux professionnels de l'IA ou aux utilisateurs finaux des informations appropriées sur les capacités et les limites du système d'IA, selon des modalités adaptées au contexte d'utilisation concerné. Ces informations pourraient comprendre le degré de précision du système d'IA, ainsi que ses limites.

5. Diversité, non-discrimination et équité

(79) Pour parvenir à une IA digne de confiance, il est nécessaire de favoriser l'inclusion et la diversité tout au long du cycle de vie du système d'IA. Outre la prise en compte et la participation de l'ensemble des parties prenantes concernées tout au long du processus, cela implique également de veiller à l'égalité d'accès au moyen de processus conçus de manière inclusive, ainsi qu'à l'égalité de traitement. Cette exigence est étroitement liée au *principe de l'équité*.

(80) ***Absence de biais injustes***. Les ensembles de données utilisés par les systèmes d'IA (tant pour leur entraînement que pour leur exploitation) peuvent être biaisés par des partis pris historiques accidentels, des omissions et des modèles de gouvernance défectueux. La persistance de ces biais pourrait être source de discrimination et de préjudice (in)directs[42] involontaires à l'encontre de certains groupes de personnes, aggravant potentiellement le préjudice et la marginalisation. Des préjudices peuvent également résulter de l'exploitation intentionnelle de préjugés (des consommateurs) ou d'une concurrence déloyale, comme l'homogénéisation des prix par le biais d'une collusion ou l'opacité d'un marché.[43] Dans la mesure du possible, les biais détectables et discriminatoires devraient être supprimés lors de la phase de collecte. La manière dont les systèmes d'IA sont mis au point (par exemple la programmation des algorithmes) peut également être entachée de biais. On peut contrer cette tendance en mettant en place des procédures de contrôle pour analyser de manière claire et transparente la finalité, les contraintes, les exigences et les décisions du système. En outre, le recrutement de personnes issues de contextes, de cultures et de disciplines différents peut garantir la diversité des opinions et devrait être encouragé.

(81) ***Accessibilité et conception universelle***. Dans le contexte des relations d'entreprise à consommateur, notamment, les systèmes devraient être centrés sur l'utilisateur et conçus de manière à permettre à toute personne d'utiliser des produits ou services d'IA, quels que soient son âge, son sexe, ses capacités ou ses caractéristiques. L'accessibilité de cette technologie aux personnes atteintes de handicaps, qui sont présentes dans tous les segments de la société, revêt une importance particulière. Les systèmes d'IA ne devraient pas adopter une approche uniforme et devraient envisager des principes de conception universelle[44] répondant aux besoins du plus large éventail possible d'utilisateurs, en suivant des normes d'accessibilité pertinentes.[45] Ce principe permettra un accès équitable et la participation active de chacun aux activités humaines informatisées existantes et émergentes, ainsi qu'aux technologies d'assistance.[46]

(82) ***Participation des parties prenantes***. Pour mettre au point des systèmes d'IA dignes de confiance, il est souhaitable de consulter les parties prenantes sur lesquelles le système est susceptible d'avoir des effets directs ou indirects tout au long de son cycle de vie. Il est bénéfique de solliciter régulièrement des commentaires, même après le déploiement, et de mettre en place des mécanismes à plus long terme de participation des parties prenantes, en veillant par exemple à l'information, la consultation et la participation des travailleurs à travers tout le processus de mise en œuvre de systèmes d'IA au sein d'organisations.

6. Bien-être sociétal et environnemental

(83) Tout comme pour les *principes de l'équité* et *de la prévention de toute atteinte*, il convient de considérer également la société au sens large, les autres êtres sensibles et l'environnement comme des parties prenantes tout au long du cycle de vie de l'IA. La durabilité et la responsabilité écologique des systèmes d'IA devraient être encouragées, et il convient de promouvoir la recherche de solutions d'IA répondant à des préoccupations

---

42 Pour une définition des formes directes et indirectes de discrimination, voir par exemple l'article 2 de la directive 2000/78/CE du Conseil du 27 novembre 2000 portant création d'un cadre général en faveur de l'égalité de traitement en matière d'emploi et de travail. Voir également l'article 21 de la charte des droits fondamentaux de l'Union européenne.

43 Voir document de l'Agence des droits fondamentaux de l'Union européenne: «BigData: Discrimination in data-supported decision making (2018)» http://fra.europa.eu/en/publication/2018/big-data-discrimination.

44 L'article 42 de la directive relative aux marchés publics prévoit que les spécifications techniques doivent prendre en compte l'accessibilité et la conception pour tous.

45 Par exemple EN 301 549.

46 Cette exigence est liée à la Convention des Nations unies relative aux droits des personnes handicapées.

de portée mondiale, par exemple les objectifs de développement durable. Idéalement, tous les êtres humains devraient bénéficier de l'IA, y compris les générations futures.

(84) ***IA durable et respectueuse de l'environnement***. Les systèmes d'IA promettent de contribuer à répondre à certaines des plus vives préoccupations de la société; il faut cependant veiller à ce que les réponses apportées soient aussi respectueuses de l'environnement que possible. Il convient, à cet égard, d'évaluer le processus de mise au point, de déploiement et d'utilisation du système, ainsi que toute sa chaîne d'approvisionnement, par exemple au moyen d'un examen critique de l'utilisation des ressources et de la consommation d'énergie au cours de l'entraînement, en réalisant les choix les moins préjudiciables. Il convient d'encourager les mesures permettant de garantir que l'ensemble de la chaîne d'approvisionnement du système d'IA respecte l'environnement.

(85) ***Incidences sociales***. L'omniprésence des systèmes d'IA sociaux[47] dans tous les domaines de notre vie (qu'il s'agisse de l'enseignement, du travail, des soins ou des loisirs) peut altérer notre conception de l'action sociale ou avoir une incidence sur nos relations et nos liens sociaux. Si les systèmes d'IA peuvent être utilisés pour renforcer les compétences sociales[48], ils peuvent également contribuer à leur détérioration. Cela pourrait également nuire au bien-être physique ou mental des personnes. Les effets de ces systèmes doivent par conséquent faire l'objet d'un contrôle et d'un examen minutieux.

(86) ***Société et démocratie.*** En plus d'évaluer l'incidence de la mise au point, du déploiement et de l'utilisation d'un système d'IA sur les individus, il convient également d'évaluer cette incidence d'un point de vue sociétal, en tenant compte de son effet sur les institutions, la démocratie et la société au sens large. L'utilisation des systèmes d'IA devrait faire l'objet d'une attention particulière dans les situations mettant en jeu le processus démocratique, non seulement la prise de décisions politiques, mais aussi les contextes électoraux.

7. Responsabilité

(87) L'exigence de la responsabilité complète les exigences susmentionnées, étroitement liées au *principe de l'équité*. Elle requiert la mise en place de mécanismes permettant de garantir l'autonomie et la responsabilité à l'égard des systèmes d'IA et de leurs résultats, tant avant qu'après leur mise en œuvre.

(88) ***Auditabilité.*** L'auditabilité implique la possibilité d'évaluer les algorithmes, les données et les processus de conception. Elle n'implique pas nécessairement que les informations sur les modèles économiques et la propriété intellectuelle en lien avec le système d'IA doivent toujours être librement accessibles. L'évaluation par des auditeurs internes et externes, ainsi que la disponibilité des rapports de ces évaluations, peuvent contribuer à la fiabilité de la technologie. Pour les applications mettant en jeu les droits fondamentaux, notamment les applications critiques pour la sécurité, les systèmes d'IA devraient pouvoir faire l'objet d'audits indépendants.

(89) ***Réduction au minimum et documentation des incidences négatives***. Il convient de garantir la capacité aussi bien de documenter les actions ou décisions contribuant à un certain résultat du système que de répondre aux conséquences d'un tel résultat. Il est particulièrement important pour les personnes touchées directement ou indirectement que les effets négatifs potentiels des systèmes d'IA soient répertoriés, analysés, documentés et réduits le plus possible. Il convient d'assurer un niveau de protection approprié aux lanceurs d'alertes, aux ONG, aux syndicats ou à d'autres entités lorsqu'ils font état de préoccupations légitimes au sujet d'un système

[47] Sont visés ici les systèmes d'IA qui communiquent et interagissent avec les êtres humains en simulant un comportement social dans les interactions entre robots et humains (IA embarquée) ou comme avatars dans la réalité virtuelle. Ce faisant, ces systèmes ont le potentiel de modifier nos pratiques socioculturelles et le tissu de notre vie sociale.

[48] Voir par exemple le projet financé par l'UE en vue de la mise au point d'un logiciel fondé sur l'IA permettant à des robots d'interagir plus efficacement avec des enfants autistes lors de sessions thérapeutiques dirigées par des êtres humains, contribuant à améliorer leurs compétences sociales et de communication: http://ec.europa.eu/research/infocentre/article_en.cfm?id=/research/headlines/news/article_19_03_12_en.html?infocentre&item=Infocentre&artid=49968.

fondé sur l'IA. Le recours aux analyses d'impact (par exemple, le «red teaming» ou certaines formes d'analyse d'impact algorithmique), tant avant que pendant la mise au point, le déploiement et l'utilisation de systèmes d'IA, peut contribuer à réduire le plus possible les effets négatifs. Ces analyses doivent être proportionnées au risque associé aux systèmes d'IA.

(90) ***Arbitrages.*** Lors de la mise en œuvre des exigences ci-dessus, des tensions pourraient survenir entre elles, ce qui pourrait rendre inévitables certains arbitrages. Ces arbitrages devraient être effectués avec raison et méthode, conformément à l'état actuel de la technique. Cela implique qu'il convient de recenser les intérêts et valeurs pertinents concernés par le système d'IA et que, en cas de conflit, les arbitrages entre eux devraient être explicitement reconnus et évalués du point de vue du risque qu'ils posent pour les principes éthiques, y compris les droits fondamentaux. Lorsqu'aucun arbitrage acceptable du point de vue éthique ne peut être déterminé, la mise au point, le déploiement et l'utilisation du système d'IA ne devraient pas se poursuivre en l'état. Toute décision concernant un arbitrage à faire devrait être raisonnée et correctement documentée. La personne chargée de prendre la décision doit être tenue responsable de la manière dont l'arbitrage pertinent est effectué et devrait en permanence reconsidérer le caractère approprié de la décision résultante, pour veiller à ce que les modifications nécessaires soient apportées au système en cas de besoin. [49]

(91) ***Recours***. Lorsqu'une incidence négative injuste se produit, il convient de prévoir des mécanismes accessibles assurant une voie de recours adéquate.[50] Savoir qu'un recours est possible lorsque les choses se passent mal est essentiel pour garantir la confiance. Il convient d'accorder une attention particulière aux personnes ou groupes vulnérables.

2. Méthodes techniques et non techniques pour parvenir à une IA digne de confiance

(92) Pour mettre en œuvre les exigences susmentionnées, des méthodes tant techniques que non techniques peuvent être appliquées. Ces méthodes englobent toutes les phases du cycle de vie d'un système d'IA. Il convient de procéder de manière continue à une évaluation des méthodes employées pour mettre en œuvre les exigences, ainsi qu'à la communication et à la justification[51] des modifications apportées aux processus de mise en œuvre. Étant donné que les systèmes d'IA évoluent et agissent de manière continue dans un environnement dynamique, la réalisation d'une IA digne de confiance est un processus continu, illustré à la figure 3 ci-dessous.

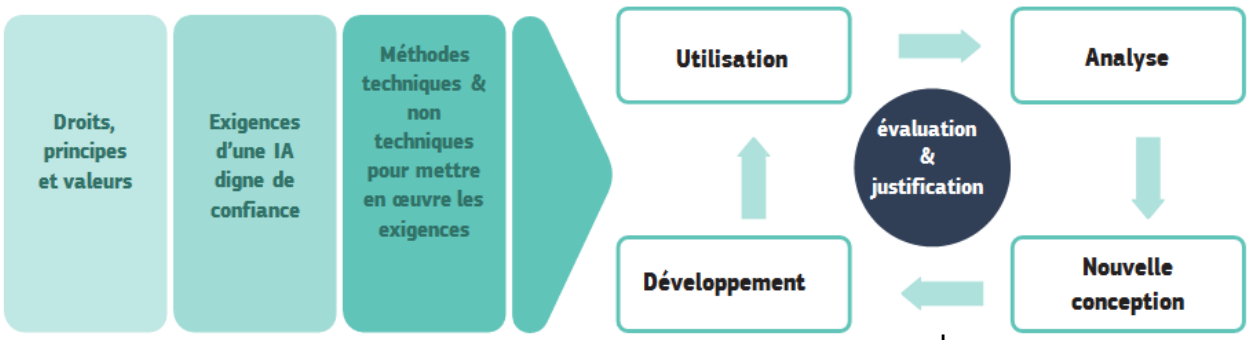


Figure 3: parvenir à une IA digne de confiance tout au long du cycle de vie du système

[49] Différents modèles de gouvernance peuvent contribuer à cet objectif. Par exemple, la présence d'un expert ou conseil éthique (et sectoriel) interne et/ou externe pourrait être utile pour mettre en évidence des domaines de conflit potentiel et proposer les manières les plus adaptées de résoudre ce conflit. Il est également utile de procéder à une consultation et à une discussion concrètes avec les parties prenantes, y compris celles susceptibles de subir les incidences négatives d'un système d'IA. Les universités européennes devraient jouer un rôle de premier plan dans la formation d'experts nécessaires dans le domaine de l'éthique.

[50] Voir également l'avis de l'Agence des droits fondamentaux de l'Union européenne sur l'amélioration de l'accès aux voies de recours dans les domaines des droits de l'homme et des entreprises au niveau de l'Union (2017), https://fra.europa.eu/en/opinion/2017/business-human-rights.

[51] Cela implique, par exemple, la justification des choix réalisés dans la conception, la mise au point et le déploiement du système pour intégrer les exigences susmentionnées.

(93) Les méthodes suivantes peuvent être considérées comme mutuellement complémentaires ou alternatives, car des exigences différentes – et des sensibilités différentes – peuvent justifier des méthodes de mise en œuvre différentes. Cet aperçu n'a pas vocation à être exhaustif ou obligatoire. Il vise plutôt à proposer une liste de méthodes possibles susceptibles de contribuer à la mise en œuvre d'une IA digne de confiance.

1. Méthodes techniques

(94) Cette section décrit les méthodes techniques pour garantir une IA digne de confiance qui puisse être incorporée aux phases de conception, de mise au point et d'utilisation d'un système d'IA. Le niveau de maturité des méthodes présentées ci-dessous varie.[52]

- *Architectures pour une IA digne de confiance*

(95) Les exigences d'une IA digne de confiance devraient être «traduites» en procédures et/ou en contraintes imposées aux procédures, qui devraient être ancrées dans l'architecture du système d'IA. Cela pourrait être accompli au moyen d'un ensemble de règles dites «listes blanches» (comportements ou états) que le système devrait toujours suivre, de restrictions dites «listes noires» relatives aux comportements ou états que le système ne devrait jamais transgresser, et de combinaisons des deux ou de garanties démontrables plus complexes concernant le comportement du système. Un processus distinct pourrait servir à contrôler le respect de ces restrictions par le système, pendant son fonctionnement.

(96) Les systèmes d'IA dotés de capacités d'apprentissage et capables d'adapter leur comportement de façon dynamique peuvent être perçus comme des systèmes non déterministes susceptibles d'afficher un comportement inattendu. Ces systèmes sont souvent considérés à travers le prisme théorique d'un cycle «sense-plan-act» (détection-planification-action). Pour que cette architecture soit adaptée à une IA digne de confiance, il convient d'intégrer les exigences à chacune des trois étapes du cycle: i) à l'étape de la «détection», le système devrait être mis au point de telle sorte qu'il reconnaisse l'ensemble des éléments présents dans l'environnement qui sont nécessaires en vue de garantir l'adhésion à ces exigences; ii) à l'étape de la «planification», le système devrait uniquement envisager des plans adhérant aux exigences, et; iii) à l'étape de l'action, les actions du système devraient être limitées aux comportements correspondant à ces exigences.

(97) L'architecture telle qu'illustrée ci-dessus est générique et ne constitue qu'une description imparfaite pour la plupart des systèmes d'IA. Elle présente toutefois des points d'ancrage pour les contraintes et les règles qui devraient être reflétées dans des modules spécifiques aux fins de la mise au point d'un système digne de confiance et perçu comme tel.

- *Éthique et état de droit dès la conception (X dès la conception)*

(98) Les méthodes destinées à garantir les valeurs dès la conception établissent des liens précis et explicites entre les principes abstraits auxquels le système doit adhérer et les décisions spécifiques de mise en œuvre. L'idée selon laquelle la conformité aux normes peut être incorporée dans la conception du système d'IA est essentielle pour cette méthode. Les entreprises ont la responsabilité de recenser les effets de leurs systèmes d'IA dès le tout début, ainsi que les normes auxquelles ces systèmes doivent se conformer pour éviter les répercussions négatives. Différentes approches «dès la conception» sont déjà largement utilisées, comme le *respect de la vie privée dès la conception* et la *sécurité dès la conception*. Comme indiqué plus haut, pour susciter la confiance, les processus, données et résultats de l'IA doivent être sûrs, et le système devrait être conçu de manière à être robuste face aux données et attaques antagonistes. Un mécanisme d'arrêt, assurant la sûreté après défaillance, devrait être mis en œuvre et le redémarrage à la suite d'un arrêt forcé (par

[52] Alors que certaines de ces méthodes sont déjà disponibles aujourd'hui, d'autres doivent encore faire l'objet de davantage de recherches. Le GEHN IA s'appuiera également sur les domaines devant faire l'objet de recherches supplémentaires aux fins de sa deuxième contribution, à savoir les recommandations en matière de politique et d'investissement.

exemple, une attaque) devrait être rendu possible.

- *Méthodes d'explication*

(99) Pour qu'un système soit digne de confiance, nous devons être en mesure de comprendre pourquoi il s'est comporté d'une certaine manière et pourquoi il a fourni une interprétation donnée. Un domaine de recherche à part entière, l'IA explicable (*Explainable AI* – XAI), essaie de répondre à cette question pour mieux comprendre les mécanismes sous-jacents du système et trouver des solutions. Il s'agit encore à ce jour d'un défi à relever pour les systèmes d'IA fondés sur des réseaux neuronaux. Les processus d'entraînement avec réseaux neuronaux peuvent déboucher sur des paramètres de réseau réglés sur des valeurs numériques difficiles à mettre en lien avec des résultats. En outre, de légères modifications apportées aux valeurs des données peuvent parfois modifier de façon spectaculaire l'interprétation, menant par exemple le système à confondre un bus scolaire avec une autruche. Cette vulnérabilité peut également être exploitée au cours d'attaques contre le système. Les méthodes recourant à l'IA explicable sont non seulement essentielles pour expliquer le comportement du système aux utilisateurs, mais également pour déployer une technologie fiable.

- *Essais et validations*

(100) Étant donné la nature non déterministe des systèmes d'IA et la mesure dans laquelle ils sont spécifiques à leurs contextes, les essais traditionnels ne sont pas suffisants. Les défaillances des concepts et des représentations utilisés par le système ne sont susceptibles de se manifester que lorsqu'un programme est appliqué à des données suffisamment réalistes. Par conséquent, pour vérifier et valider le traitement des données, le modèle sous-jacent doit faire l'objet d'un contrôle attentif tant au cours de l'entraînement que du déploiement en ce qui concerne sa stabilité, sa robustesse et son fonctionnement dans des limites bien comprises et prévisibles. Il convient de veiller à ce que le résultat du processus de planification corresponde aux données d'entrée, et que les décisions soient prises d'une façon qui permette la validation du processus sous-jacent.

(101) Les essais et la validation du système devraient avoir lieu le plus tôt possible, pour veiller à ce que le système se comporte de la manière prévue tout au long de son cycle de vie et notamment après son déploiement. Ils devraient porter sur l'ensemble des éléments d'un système d'IA, y compris les données, les modèles pré-entraînés, les environnements et le comportement du système dans son ensemble, et être conçus et mis en œuvre par un groupe de personnes le plus divers possible. Plusieurs indicateurs devraient être définis pour couvrir les catégories faisant l'objet d'essais dans différentes perspectives. Des essais antagonistes réalisés par des «équipes rouges» de confiance et diversifiées, tentant délibérément de «pénétrer» le système à la recherche de failles, ainsi que des «primes au bogue» incitant les utilisateurs externes à détecter et signaler de manière responsable les erreurs et les faiblesses du système, peuvent être envisagés. Enfin, il convient de veiller à ce que les résultats ou actions correspondent aux résultats des processus préalables, en les comparant aux règles définies antérieurement pour faire en sorte que celles-ci ne soient pas violées.

- *Qualité des indicateurs de service*

(102) Un niveau de qualité approprié des indicateurs de service peut être défini pour les systèmes d'IA, afin de faire en sorte qu'un point de comparaison existe pour déterminer s'ils ont été testés et mis au point en tenant compte de la sécurité et de la sûreté. Ces indicateurs pourraient comprendre des mesures pour évaluer les essais et l'entraînement des algorithmes ainsi que des indicateurs logiciels traditionnels de la fonctionnalité, de la performance, de la facilité d'utilisation, de la fiabilité, de la sécurité et de la maintenabilité.

2. Méthodes non techniques

(103) Cette section décrit un éventail de méthodes non techniques susceptibles de jouer un rôle important pour obtenir et préserver une IA digne de confiance. Ces méthodes devraient également faire l'objet d'une **évaluation constante**.

- *Réglementation*

(104) Comme indiqué plus haut, une réglementation est déjà en place pour soutenir la fiabilité de l'IA, comme la législation relative à la sécurité des produits et les cadres applicables en matière de responsabilité. Dans la mesure où nous considérons qu'il pourrait être nécessaire de réviser ou d'adapter la réglementation, ou d'en adopter de nouvelles – pour servir tant de garanties que de catalyseurs – cet aspect sera abordé dans le cadre de notre deuxième contribution, qui consistera à formuler des recommandations en matière de politique et d'investissement.

- *Codes de conduite*

(105) Les organisations et les parties prenantes peuvent adopter les lignes directrices et adapter leurs chartes de responsabilité de l'organisation, leurs indicateurs de performances clés («KPI»), leurs codes de conduite ou règles internes pour y ajouter l'objectif de parvenir à une IA digne de confiance. Une organisation travaillant à la mise au point d'un système d'IA peut, de manière plus générale, documenter ses intentions, ainsi que les appuyer sur des normes relatives à certaines valeurs souhaitables, telles que les droits fondamentaux, la transparence et la prévention des préjudices.

- *Normalisation*

(106) Les normes, en matière par exemple de conception, de fabrication et de pratiques commerciales, peuvent fonctionner en tant que système de gestion de la qualité pour les utilisateurs de l'IA, consommateurs, organisations, instituts de recherche et pouvoirs publics, en offrant la possibilité de reconnaître et d'encourager un comportement éthique par leurs décisions d'achats. Outre les normes conventionnelles, il existe des approches de corégulation: systèmes d'agrément, codes de déontologie professionnels ou normes relatives à une conception conforme aux droits fondamentaux. On compte notamment parmi les exemples actuels les normes ISO ou les séries de normes IEEE P7000. Toutefois, un futur label «IA digne de confiance» pourrait être approprié, qui confirmerait par référence à des normes techniques spécifiques que le système est conforme, par exemple, en matière de sûreté, de robustesse technique et d'explicabilité.

- *Certification*

(107) Si on ne peut pas s'attendre à ce que tout le monde soit capable de comprendre totalement le fonctionnement et les effets des systèmes d'IA, on pourrait imaginer des organisations qui soient en mesure d'attester auprès du grand public qu'un système d'IA est transparent, responsable et juste.[53] Ces certifications appliqueraient des normes définies pour différents domaines d'application et techniques d'IA, dûment alignées sur les normes industrielles et sociétales des différents contextes. Une certification ne pourra toutefois jamais remplacer la responsabilité. Elle devrait par conséquent s'accompagner de cadres de responsabilité, y compris de clauses de non-responsabilité ainsi que de mécanismes de révision et de correction[54].

- *La responsabilité au moyen de cadres de gouvernance*

(108) Les organisations devraient définir des cadres de gouvernance, tant internes qu'externes, garantissant la responsabilité à l'égard des dimensions éthiques des décisions associées à la mise au point, au déploiement et à l'utilisation de l'IA. Cela pourrait, par exemple, comprendre la nomination d'une personne chargée des questions d'éthique en lien avec l'IA, ou d'un groupe ou conseil interne/externe traitant de ces questions. Ces personnes, groupes ou conseils pourraient être chargés d'assurer une supervision et de formuler des conseils. Comme indiqué plus haut, des spécifications et/ou organismes de certification peuvent jouer un rôle à cet

[53] Comme le préconise par exemple l'IEEE dans son initiative relative à une conception alignée sur le plan éthique: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

[54] Pour plus d'informations sur les limites de la certification, voir: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

effet. Des canaux de communication devraient être mis en place avec des groupes de supervision issus des secteurs public et/ou privé, pour partager les bonnes pratiques, discuter des dilemmes ou signaler des problèmes émergents liés à des préoccupations éthiques. De tels mécanismes peuvent compléter mais pas remplacer le contrôle juridique (par exemple, via la nomination d'un responsable de la protection des données ou des mesures équivalentes, qui sont juridiquement requises par la législation sur la protection des données).

- *Éducation et sensibilisation pour encourager un état d'esprit éthique*

(109) Une IA digne de confiance encourage la participation informée de toutes les parties prenantes. La communication, l'éducation et la formation jouent un rôle important, tant pour veiller à la diffusion des connaissances sur les incidences potentielles des systèmes d'IA que pour informer la population qu'elle peut participer à l'orientation du développement de la société. Cela concerne l'ensemble des parties prenantes, par exemple celles qui sont impliquées dans la création de produits (concepteurs et développeurs), les utilisateurs (entreprises ou personnes) et d'autres groupes concernés (ceux qui n'achèteront ou n'utiliseront pas nécessairement un système d'IA mais au nom de qui des décisions sont prises par un système d'IA, et la société au sens large). L'acquisition de connaissances de base en matière d'IA devrait être encouragée dans toute la société. Une condition préalable pour éduquer le public est de veiller à ce que les éthiciens possèdent les compétences et la formation requises dans ce domaine.

- *Participation des parties prenantes et dialogue social*

(110) Les avantages de l'IA sont nombreux, et l'Europe doit veiller à ce qu'ils soient à la disposition de chacun. Cela nécessite une discussion ouverte et la participation des partenaires sociaux, des parties prenantes ainsi que du grand public. De nombreuses organisations s'appuient déjà sur des groupes de parties prenantes pour discuter de l'utilisation des systèmes d'IA et de l'analyse des données. Ces groupes sont composés de différents membres, tels que des experts juridiques, des experts techniques, des éthiciens, des représentants de consommateurs et des travailleurs. La recherche active d'une participation et d'un dialogue concernant l'utilisation et les incidences des systèmes d'IA contribue à l'évaluation des résultats et des approches, et peut s'avérer particulièrement utile dans les cas complexes.

- *Diversité et équipes de conception inclusives*

(111) La diversité et l'inclusion jouent un rôle essentiel dans la mise au point de systèmes d'IA destinés à être utilisés dans le monde réel. Alors que les systèmes d'IA réalisent davantage de tâches de manière autonome, il est essentiel que les équipes qui conçoivent, mettent au point, testent, entretiennent, déploient et/ou achètent ces systèmes reflètent la diversité des utilisateurs et de la société en général. Cela contribue à l'objectivité et à la prise en compte de différents points de vue, besoins et objectifs. Idéalement, les équipes doivent non seulement être diversifiées en ce qui concerne le genre, la culture et l'âge, mais également du point de vue du parcours professionnel et des compétences.

Orientations essentielles dérivées du chapitre II:

- ✓ Veiller à ce que l'ensemble du cycle de vie du système d'IA réponde aux exigences d'une IA digne de confiance: 1) action humaine et contrôle humain, 2) robustesse technique et sécurité, 3) respect de la vie privée et gouvernance des données, 4) transparence, 5) diversité, non-discrimination et équité, 6) bien-être sociétal et environnemental, et 7) responsabilité.
- ✓ Envisager des méthodes techniques et non techniques afin de garantir la mise en œuvre de ces exigences.
- ✓ Encourager la recherche et l'innovation en vue de contribuer à l'évaluation des systèmes d'IA et de soutenir la mise en œuvre des exigences; diffuser les résultats et adresser les questions au grand public, et veiller à ce qu'une formation dans le domaine l'éthique en matière d'IA soit systématiquement dispensée à la nouvelle génération d'experts.

- ✓ Fournir, clairement et de façon proactive, des informations aux parties prenantes sur les capacités et les limites des systèmes d'IA, afin de leur permettre de formuler des attentes réalistes, ainsi que sur la manière dont les exigences sont mises en œuvre. Faire preuve de transparence sur le fait qu'elles interagissent avec un système d'IA.
- ✓ Faciliter la traçabilité et l'auditabilité des systèmes d'IA, en particulier dans les contextes et situations critiques.
- ✓ Mobiliser les parties prenantes tout au long du cycle de vie des systèmes d'IA. Encourager la formation et l'éducation afin que toutes les parties prenantes soient renseignées sur l'IA digne de confiance et formées dans ce domaine.
- ✓ Avoir conscience qu'il peut exister des tensions fondamentales entre différents principes et exigences. Recenser, évaluer, documenter et communiquer de manière continue ces arbitrages et leurs solutions.

III. Chapitre III: évaluation d'une IA digne de confiance

(112) Sur la base des exigences essentielles du chapitre II, ce chapitre établit une **liste d'évaluation** non exhaustive **pour une IA digne de confiance** (version pilote) permettant de **concrétiser une IA digne de confiance.** Cette liste s'applique notamment aux systèmes d'IA qui interagissent directement avec les utilisateurs et est avant tout destinée aux développeurs et aux prestataires chargés du déploiement de systèmes d'IA (qu'ils aient été mis au point en interne ou obtenus auprès de tiers). Elle ne porte pas sur la concrétisation du premier élément caractéristique d'une IA digne de confiance (la licéité). Cette liste d'évaluation, dont le respect ne constitue pas une preuve de conformité à la législation, n'a pas vocation à fournir des orientations visant à garantir le respect de la législation applicable. Vu la spécificité des applications propres aux systèmes d'IA, cette liste d'évaluation devra être adaptée aux cas d'utilisation et contextes spécifiques dans lesquels le système fonctionne. En outre, ce chapitre propose une recommandation générale sur la manière de mettre en œuvre la liste d'évaluation pour une IA digne de confiance au moyen d'une structure de gouvernance englobant tant le niveau opérationnel que le niveau de l'encadrement.

(113) La liste d'évaluation et la structure de gouvernance seront élaborées en étroite collaboration avec les parties prenantes des secteurs public et privé. Ce processus sera mené en tant que processus «pilote» et permettra de recueillir de nombreuses réactions dans le cadre de deux processus parallèles:

a. un processus qualitatif garantissant la représentation, auquel un nombre limité d'entreprises, d'organisations et d'institutions (de différents secteurs et de différentes tailles) adhérera pour tester la liste d'évaluation et la structure de gouvernance dans la pratique et fournir un retour d'information détaillé;

b. un processus quantitatif auquel toutes les parties prenantes pourront adhérer pour tester la liste d'évaluation et formuler des commentaires au moyen d'une consultation ouverte.

(114) À la suite de la phase pilote, nous intégrerons les résultats de ces processus à la liste d'évaluation et préparerons une version révisée début 2020. L'objectif est d'obtenir un cadre pouvant être appliqué de manière transversale à l'ensemble des applications et donc de jeter les bases d'une IA digne de confiance dans tous les domaines. Une fois ces bases établies, un cadre sectoriel ou spécifique aux applications pourrait être défini.

Gouvernance

(115) Des entreprises, organisations et institutions pourraient s'intéresser à la manière de mettre en œuvre la liste d'évaluation pour une IA digne de confiance au sein de leur structure. Pour ce faire, elles pourraient inclure le processus d'évaluation à leurs mécanismes de gouvernance existants, ou mettre en œuvre de

nouveaux processus. Ce choix dépendra de la structure interne de l'organisation ainsi que de sa taille et des ressources dont elle dispose.

(116) Des recherches[55] indiquent que tout changement requiert nécessairement l'attention des cadres supérieurs. Elles indiquent aussi que mobiliser l'ensemble des parties prenantes d'une entreprise, organisation ou institution accroît l'acceptation et la pertinence de l'introduction d'un nouveau processus (technologique ou non)[56]. Nous recommandons par conséquent la mise en œuvre d'un processus comprenant la participation tant au niveau opérationnel qu'au niveau des cadres supérieurs.

Niveau	Rôles pertinents (en fonction de l'organisation)
Encadrement et organe supérieur	L'encadrement supérieur étudie et évalue la mise au point, le déploiement ou l'achat de l'IA, en tant qu'échelon supérieur pour l'évaluation de l'ensemble des innovations et utilisations de l'IA, lorsque des préoccupations majeures sont détectées. Il y associe les personnes concernées par l'éventuelle introduction de systèmes d'IA (par exemple, des travailleurs) et leurs représentants tout au long du processus via des procédures d'information, de consultation et de participation.
Service chargé des questions de conformité, de licéité et de responsabilité de l'organisation	Ce service contrôle l'utilisation de la liste d'évaluation et sa nécessaire évolution pour répondre aux changements technologiques ou réglementaires. Il met à jour les normes ou règles internes relatives aux systèmes d'IA et veille à ce que l'utilisation de ces systèmes respecte le cadre juridique et réglementaire en vigueur et les valeurs de l'organisation.
Service chargé de la mise au point des produits et services ou équivalent	Le service chargé de la mise au point des produits et services utilise la liste d'évaluation pour évaluer les produits et services fondés sur l'IA et assure la journalisation de tous les résultats. Ces résultats font l'objet de discussions au niveau de l'encadrement, qui approuve en fin de compte les applications nouvelles ou révisées fondées sur l'IA.
Assurance qualité	Le service d'assurance qualité (ou équivalent) contrôle les résultats de la liste d'évaluation et prend des mesures pour faire remonter le problème à un échelon supérieur lorsque le résultat n'est pas satisfaisant ou que des résultats imprévus sont détectés.
RH	Le service RH veille à ce que les développeurs de systèmes d'IA possèdent un bon éventail de compétences et présentent une diversité de profils. Il fait en sorte qu'un niveau approprié de formation soit dispensé au sein de l'organisation au sujet de l'IA digne de confiance.
Achats	Le service des achats veille à ce que la procédure d'achat de produits ou services dotés fondés sur l'IA prévoie un contrôle de leur fiabilité.
Activités quotidiennes	Les développeurs et gestionnaires de projets utilisent la liste d'évaluation dans leur travail quotidien et documentent les résultats et les conséquences de l'évaluation.

[55] https://www.mckinsey.com/business-functions/operations/our-insights/secrets-of-successful-change-implementation.

[56] Voir par exemple A. Bryson, E. Barth et H. Dale-Olsen, T*he Effects of Organisational change on worker well-being and the moderating role of trade unions*, *ILRReview*, 66(4), juillet 2013; Jirjahn, U. et Smith, S.C. (2006). *What Factors Lead Management to Support or Oppose Employee Participation—With and Without Works Councils? Hypotheses and Evidence from Germany's Industrial Relations*, 45(4), 650–680; Michie, J. et Sheehan, M. (2003). *Labour market deregulation, "flexibility" and innovation*, *Cambridge Journal of Economics*, 27(1), 123–143.

Utilisation de la liste d'évaluation pour une IA digne de confiance

(117) Lorsque la liste d'évaluation est utilisée en pratique, nous recommandons de prêter attention non seulement aux sujets de préoccupation, mais également aux questions auxquelles aucune réponse ne peut (facilement) être apportée. Le manque de diversité dans les aptitudes et les compétences de l'équipe chargée de la mise au point et des essais du système d'IA pourrait être un problème potentiel, et il pourrait par conséquent être nécessaire de mobiliser d'autres parties prenantes à l'intérieur ou à l'extérieur de l'organisation. Il est fortement recommandé de consigner tous les résultats d'un point de vue tant technique que managérial, en veillant à ce que la résolution du problème puisse être comprise à tous les niveaux de la structure de gouvernance.

(118) La liste d'évaluation a vocation à guider les professionnels de l'IA dans la mise au point, le déploiement et l'utilisation d'une IA digne de confiance. L'évaluation devrait être adaptée au cas d'utilisation spécifique de manière proportionnée. Au cours de la phase pilote, des domaines sensibles spécifiques pourraient être révélés et le besoin d'autres dispositions particulières dans ce genre de cas sera évalué à la prochaine étape. Si cette liste d'évaluation n'apporte pas de réponses concrètes aux questions soulevées, elle encourage la réflexion sur les démarches susceptibles de contribuer à la fiabilité des systèmes d'IA et sur les démarches potentielles à adopter à cet égard.

Lien avec la législation et les processus existants

(119) Il est également important pour les personnes participant à la mise au point, au déploiement et à l'utilisation de l'IA de reconnaître que différentes législations existantes, portant sur l'application de processus spécifiques et l'interdiction de résultats particuliers, pourraient se chevaucher et coïncider avec certaines des mesures figurant dans la liste d'évaluation. Par exemple, la législation sur la protection des données définit un ensemble d'exigences juridiques que les personnes mobilisées pour la collecte et le traitement de données à caractère personnel doivent appliquer. Pourtant, étant donné qu'une IA digne de confiance passe aussi par un traitement éthique des données, les procédures et règles internes visant à assurer la conformité avec la législation sur la protection des données pourraient également contribuer à faciliter le traitement éthique des données et peuvent donc compléter les processus juridiques existants. Cette liste d'évaluation, dont le respect *ne constitue pas* une preuve de conformité à la législation, n'a pourtant pas vocation à fournir des orientations visant à garantir le respect de la législation applicable. Elle vise plutôt à présenter un ensemble de questions spécifiques aux destinataires, pour veiller à ce que leur approche de la mise au point et du déploiement de l'IA soit orientée vers l'obtention d'une IA digne de confiance.

(120) De la même manière, de nombreux professionnels de l'IA disposent déjà d'outils d'évaluation et de processus de développement de logiciels pour veiller également à la conformité avec des normes non juridiques. L'évaluation ci-dessous ne devrait pas nécessairement être réalisée de manière isolée, mais peut être incorporée à de telles pratiques existantes.

LISTE D'EVALUATION POUR UNE IA DIGNE DE CONFIANCE (VERSION PILOTE)

1. Action humaine et contrôle humain

Droits fondamentaux:

✓ Dans les cas d'utilisation susceptibles d'entraîner des effets négatifs sur les droits fondamentaux,

avez-vous réalisé une analyse d'impact sur les droits fondamentaux? Avez-vous déterminé et documenté le recours potentiel à des arbitrages entre les différents principes et droits?

- ✓ Le système d'IA interagit-il avec la prise de décision par un utilisateur final humain (par exemple, en recommandant des mesures ou décisions à prendre, ou en présentant des choix possibles)?
  - Dans de tels cas, existe-t-il un risque que le système d'IA affecte l'autonomie humaine en interférant de manière involontaire avec le processus décisionnel de l'utilisateur final?
  - Estimez-vous qu'un système d'IA devrait communiquer aux utilisateurs qu'une décision, un contenu, un conseil ou un résultat découlent d'une décision algorithmique?
  - Lorsque le système d'IA comporte un robot ou système conversationnel, les utilisateurs humains sont-ils informés du fait qu'ils interagissent avec un agent virtuel?

Action humaine:

- ✓ Lorsque le système d'IA est intégré dans un processus de travail, avez-vous réfléchi à la répartition des tâches entre le système d'IA et les travailleurs humains pour permettre des interactions constructives ainsi qu'une supervision et un contrôle humains appropriés?
  - Le système d'IA renforce-t-il ou augmente-t-il les capacités humaines?
  - Avez-vous prévu des garanties pour empêcher toute confiance ou dépendance excessives envers le système d'IA dans les processus de travail?

Contrôle humain:

- ✓ Avez-vous réfléchi au niveau approprié de contrôle humain pour le système d'IA et le cas d'utilisation en question?
  - Pouvez-vous décrire le niveau de contrôle ou de participation humains, le cas échéant? Qui est «l'humain aux manettes» et à quel moment y a-t-il intervention humaine, ou avec quels outils?
  - Avez-vous mis en place des mécanismes et des mesures pour garantir un contrôle ou une supervision humains potentiels de cette nature, ou pour veiller à ce que les décisions soient prises sous la responsabilité globale d'êtres humains?
  - Avez-vous pris des mesures pour permettre la réalisation d'audits et résoudre des questions liées à la gouvernance de l'autonomie de l'IA?
- ✓ Dans le cas d'un système d'IA ou d'une utilisation capables d'autoapprentissage ou autonomes, avez-vous mis en place des mécanismes plus spécifiques de contrôle et de supervision?
  - Quel type de mécanismes de détection et de réponse avez-vous mis sur pied pour évaluer le risque que des problèmes surviennent?
  - Avez-vous veillé à la présence d'un «bouton d'arrêt» ou à l'existence d'une procédure pour suspendre, en cas de besoin, une opération en toute sécurité? Cette procédure suspend-elle le processus dans sa totalité, en partie, ou délègue-t-elle le contrôle à un être humain?

2. Robustesse technique et sécurité

Résilience aux attaques et sécurité:

- ✓ Avez-vous évalué des formes d'attaques potentielles auxquelles le système d'IA pourrait être vulnérable?
  - Avez-vous en particulier envisagé différents types et différentes natures de vulnérabilités, comme la pollution des données, l'infrastructure physique, les cyberattaques?
- ✓ Avez-vous prévu des mesures ou systèmes pour veiller à l'intégrité et à la résilience du système d'IA face à de potentielles attaques?
- ✓ Avez-vous évalué le comportement de votre système dans des situations ou des environnements imprévus?
- ✓ Avez-vous envisagé si, et dans quelle mesure, votre système pourrait avoir un double usage? Le cas échéant, avez-vous pris des mesures préventives appropriées contre un tel cas de figure (y compris, par exemple, ne pas publier la recherche ou ne pas déployer le système)?

Solutions de secours et sécurité générale:

- ✓ Avez-vous veillé à ce que votre système dispose de suffisamment de solutions de secours pour faire face à d'éventuelles attaques antagonistes ou autres situations imprévues (par exemple, procédures de relais technique ou demande de communication avec un opérateur humain avant d'agir)?
- ✓ Avez-vous envisagé le niveau de risque posé par le système d'IA dans ce cas d'utilisation spécifique?
  - Avez-vous mis en place un processus pour mesurer et évaluer les risques et la sécurité?
  - Avez-vous fourni les informations nécessaires en cas de risque pour l'intégrité physique humaine?
  - Avez-vous réfléchi à une politique d'assurance pour couvrir les dégâts potentiels provoqués par le système d'IA?
  - Avez-vous recensé les risques potentiels en matière de sécurité d'(autres) utilisations prévisibles de la technologie, y compris d'utilisation abusive accidentelle ou malveillante? Existe-t-il un plan pour atténuer ou gérer ces risques?
- ✓ Avez-vous évalué s'il est probable que le système d'IA cause des dommages ou préjudices aux utilisateurs ou à des tiers? Le cas échéant, avez-vous évalué la probabilité, les dommages potentiels, le public concerné et la gravité?
  - En cas de risque qu'un système d'IA cause des dommages, avez-vous réfléchi à des règles de responsabilité et de protection des consommateurs, et de quelle manière en avez-vous tenu compte?
  - Avez-vous réfléchi à l'incidence potentielle ou au risque en matière de sécurité sur l'environnement ou les animaux?
  - Vous êtes-vous demandé, dans le cadre de votre analyse des risques, si des problèmes de sécurité ou de réseau (par exemple, des menaces pesant sur la cybersécurité) pourraient mettre en péril la sécurité ou entraîner des préjudices du fait d'un comportement involontaire du

système d'IA?

- ✓ Avez-vous évalué l'incidence probable d'une défaillance de votre système d'IA entraînant la production de résultats erronés, l'indisponibilité de votre système, ou la production de résultats inacceptables pour la société (par exemple, pratiques discriminatoires)?
  - Avez-vous mis en place des seuils et une gouvernance pour les scénarios ci-dessus afin de déclencher d'autres plans/solutions de secours?
  - Avez-vous défini et testé des solutions de secours?

Précision

- ✓ Avez-vous évalué le niveau de précision et la définition de la précision nécessaires dans le contexte du système d'IA et du cas d'utilisation concerné?
  - Avez-vous réfléchi à la manière dont la précision est mesurée et assurée?
  - Avez-vous mis en place des mesures pour veiller à ce que les données utilisées soient exhaustives et à jour?
  - Avez-vous mis en place des mesures pour évaluer si des données supplémentaires sont nécessaires, par exemple pour améliorer la précision et éliminer les biais?
- ✓ Avez-vous évalué le préjudice que causeraient des prédictions inexactes du système d'IA?
- ✓ Avez-vous prévu des moyens de mesurer si votre système produit un nombre inacceptable de prédictions inexactes?
- ✓ En cas de prédictions inexactes, avez-vous mis en place une série d'étapes pour résoudre le problème?

Fiabilité et reproductibilité:

- ✓ Avez-vous mis en place une stratégie afin de contrôler le système d'IA et de vous assurer qu'il répond aux objectifs, aux finalités et aux applications prévues?
  - Avez-vous vérifié si des contextes spécifiques ou conditions particulières doivent être pris en compte pour garantir la reproductibilité?
  - Avez-vous mis en place des processus ou méthodes de vérification pour mesurer et garantir les différents aspects de la fiabilité et de la reproductibilité?
  - Avez-vous mis en place des processus visant à décrire certains réglages susceptibles d'entraîner une défaillance du système d'IA?
  - Avez-vous clairement documenté et appliqué ces processus aux fins des essais et de la vérification de la fiabilité du système d'IA?

    Avez-vous mis en place un mécanisme ou une communication pour garantir aux utilisateurs (finaux) la fiabilité du système d'IA?

3. <u>Respect de la vie privée et gouvernance des données</u>

Respect de la vie privée et protection des données:

- ✓ En fonction du cas d'utilisation, avez-vous mis sur pied un mécanisme permettant à autrui de signaler des problèmes en rapport avec le respect de la vie privée et la protection des données durant les processus suivis par le système d'IA pour la collecte des données (aux fins de l'entraînement et du fonctionnement) et le traitement des données?
- ✓ Avez-vous évalué le type et la portée des données constituant vos ensembles de données (par exemple, si elles contiennent des données à caractère personnel)?
- ✓ Avez-vous réfléchi à des manières de mettre au point le système d'IA ou d'entraîner le modèle sans utiliser (ou en utilisant de manière limitée) des données potentiellement sensibles ou à caractère personnel?
- ✓ Avez-vous intégré des mécanismes de notification et de contrôle concernant les données à caractère personnel en fonction du cas d'utilisation (comme un consentement valable et la possibilité de révoquer le consentement, le cas échéant)?
- ✓ Avez-vous pris des mesures pour renforcer le respect de la vie privée, par exemple des mesures de cryptage, d'anonymisation et d'agrégation?
- ✓ Lorsqu'il existe un responsable de la protection des données, avez-vous mobilisé cette personne à un stade précoce dans le processus?

Qualité et intégrité des données:

- ✓ Avez-vous aligné votre système sur d'éventuelles normes pertinentes (par exemple, ISO, IEEE) ou des protocoles largement adoptés dans le cadre de votre gestion et de votre gouvernance quotidiennes des données?
- ✓ Avez-vous mis sur pied des mécanismes de contrôle pour la collecte, le stockage, le traitement et l'utilisation des données?
- ✓ Avez-vous évalué la mesure dans laquelle vous contrôlez la qualité des sources externes des données utilisées?
- ✓ Avez-vous mis en place des processus pour garantir la qualité et l'intégrité de vos données? Avez-vous envisagé d'autres processus? De quelle manière vérifiez-vous que vos ensembles de données n'ont pas été compromis ou piratés?

Accès aux données:

- ✓ Quels protocoles, processus et procédures ont été suivis pour gérer et garantir la gouvernance appropriée des données?
  - Avez-vous analysé qui peut accéder aux données des utilisateurs et dans quelles circonstances?
  - Avez-vous veillé à ce que ces personnes soient qualifiées, qu'elles aient effectivement besoin d'accéder aux données et à ce qu'elles disposent des compétences nécessaires pour comprendre précisément la politique de protection des données?
  - Avez-vous prévu un mécanisme de contrôle pour consigner quand, où, comment, par qui et dans quel but les données ont été consultées?

4. Transparence

Traçabilité:

- ✓ Avez-vous mis des mesures en place susceptibles de garantir la traçabilité? Cela pourrait consister à documenter:
  - ➢ Les méthodes appliquées aux fins de la conception et de la mise au point du système algorithmique:
    - o dans le cas d'un système d'IA fondé sur des règles, la méthode de programmation ou la manière dont le modèle a été mis au point devraient être documentées;
    - o dans le cas d'un système d'IA fondé sur l'apprentissage, la méthode d'entraînement de l'algorithme, y compris quelles données d'entrée ont été collectées et sélectionnées, et dans quelles conditions, devrait être documentée.
  - ➢ Les méthodes appliquées pour tester et valider le système algorithmique:
    - o dans le cas d'un système d'IA fondé sur des règles, les scénarios ou cas utilisés pour tester et valider devraient être documentés;
    - o dans le cas d'un système d'IA fondé sur l'apprentissage, les informations relatives aux données utilisées pour tester et valider devraient être documentées.
  - ➢ Les résultats du système algorithmique:
    - o les résultats d'un algorithme ou les décisions qu'il prend, ainsi que les éventuelles autres décisions qui résulteraient de différents cas (par exemple, pour d'autres sous-groupes d'utilisateurs) devraient être documentés.

Explicabilité:

- ✓ Avez-vous évalué la mesure dans laquelle les décisions prises, et donc les résultats obtenus, par le système d'IA peuvent être compris?
- ✓ Avez-vous veillé à ce qu'une explication de la raison pour laquelle un système a procédé à un certain choix entraînant un certain résultat puisse être rendue compréhensible pour l'ensemble des utilisateurs qui pourraient souhaiter obtenir une explication?
- ✓ Avez-vous évalué la mesure dans laquelle la décision du système influence les processus décisionnels de l'organisation?
- ✓ Avez-vous évalué pourquoi ce système particulier a été déployé dans ce domaine spécifique?
- ✓ Avez-vous évalué le modèle économique concernant ce système (par exemple, en quoi crée-t-il de la valeur pour l'organisation)?
- ✓ Avez-vous conçu le système d'IA en ayant dès le départ l'interprétation à l'esprit?
  - ▪ Avez-vous cherché à utiliser le modèle le plus simple et le plus facile à interpréter pour l'application en question?
  - ▪ Avez-vous évalué si vous êtes en mesure d'analyser les données que vous avez utilisées aux fins de l'entraînement et des essais? Cela peut-il être modifié et actualisé au fil du temps?

- Avez-vous évalué si des solutions s'offrent à vous suite à l'entraînement et à la mise au point du modèle pour examiner l'interprétation ou si vous avez accès à la séquence des opérations du modèle?

Communication:

✓ Avez-vous informé les utilisateurs (finaux) – au moyen d'une clause de non-responsabilité ou de tout autre moyen – qu'ils interagissent avec un système d'IA et pas avec un autre être humain? Avez-vous indiqué clairement que votre système est doté de l'IA?

✓ Avez-vous mis en place des mécanismes pour informer les utilisateurs des raisons et critères expliquant les résultats du système d'IA?

- Les utilisateurs visés en sont-ils informés de manière claire et intelligible?
- Avez-vous établi des processus pour tenir compte des commentaires des utilisateurs et utiliser ces commentaires pour adapter le système?
- Avez-vous également communiqué les risques potentiels ou perçus, tels que les biais?
- En fonction du cas d'utilisation, avez-vous également réfléchi à la communication et à la transparence envers d'autres publics, des tiers ou le grand public?

✓ Avez-vous clairement indiqué la finalité du système d'IA et qui ou ce qui pourrait bénéficier du produit/service?

- Les scénarios d'utilisation du produit ont-ils été définis et clairement expliqués, en envisageant également d'autres moyens de communication pour veiller à ce qu'ils soient compréhensibles et appropriés pour le destinataire?
- En fonction du cas d'utilisation, avez-vous réfléchi à la psychologie humaine et aux potentielles limites humaines, comme le risque de confusion, les biais de confirmation ou la fatigue cognitive?

✓ Avez-vous clairement expliqué les caractéristiques, les limites et les éventuelles lacunes du système d'IA:

- s'agissant de la mise au point: à toute personne chargée de son déploiement pour en faire un produit ou service?
- s'agissant du déploiement: à l'utilisateur final ou au consommateur?

5. <u>Diversité, non-discrimination et équité</u>

Éviter les biais injustes:

✓ Avez-vous prévu une stratégie ou un ensemble de procédures pour éviter de créer ou de renforcer des biais injustes dans le système d'IA, en ce qui concerne tant l'utilisation des données d'entrée que la conception de l'algorithme?

- Avez-vous évalué et reconnu les éventuelles limites provenant de la composition des ensembles de données utilisés?
- Avez-vous réfléchi à la diversité et à la représentativité des utilisateurs dans les données? Avez-

vous procédé à des essais portant sur des populations spécifiques ou des cas d'utilisation problématiques?

- Avez-vous recherché et utilisé les outils techniques disponibles pour améliorer votre compréhension des données, du modèle et de la performance?
- Avez-vous mis en place des processus pour tester et contrôler les biais éventuels au cours de la phase de mise au point, de déploiement et d'utilisation du système?

✓ En fonction du cas d'utilisation, avez-vous prévu un mécanisme permettant à autrui de signaler des problèmes liés aux biais, à la discrimination ou aux mauvaises performances du système d'IA?

- Avez-vous envisagé des mesures et des moyens de communication clairs pour savoir comment et à qui ces problèmes peuvent être signalés?
- Avez-vous tenu compte non seulement des utilisateurs (finaux) mais également des autres personnes susceptibles d'être indirectement affectées par le système d'IA?

✓ Avez-vous évalué si, dans des conditions identiques, une éventuelle variabilité des décisions est possible?

- Le cas échéant, avez-vous réfléchi aux causes probables?
- Concernant la variabilité, avez-vous mis sur pied un mécanisme de mesure ou d'évaluation de l'incidence potentielle de cette variabilité sur les droits fondamentaux?

✓ Avez-vous prévu une définition appropriée de l'«équité» que vous appliquez dans la conception des systèmes d'IA?

- Votre définition est-elle couramment utilisée? Avez-vous envisagé d'autres définitions avant de choisir celle-ci?
- Avez-vous prévu une analyse quantitative ou des indicateurs pour mesurer et tester la définition appliquée de l'équité?
- Avez-vous mis sur pied des mécanismes visant à garantir l'équité dans vos systèmes d'IA? Avez-vous envisagé d'autres mécanismes potentiels?

Accessibilité et conception universelle:

✓ Avez-vous veillé à ce que le système d'IA réponde aux besoins d'un large ensemble de préférences et de capacités individuelles?

- Avez-vous évalué si le système d'IA peut être utilisé par les personnes présentant des besoins spécifiques ou un handicap ou qui sont exposées au risque d'exclusion? Comment cet aspect a-t-il été intégré à la conception du système et comment est-il vérifié?
- Avez-vous veillé à ce que les informations relatives au système d'IA soient également accessibles aux utilisateurs de technologies d'assistance?
- Avez-vous mobilisé ou consulté cette communauté d'utilisateurs au cours de la phase de mise au point du système d'IA?

✓ Avez-vous tenu compte de l'incidence de votre système d'IA sur le groupe d'utilisateurs potentiels?

- L'équipe participant à la mise au point du système d'IA est-elle représentative de votre groupe cible d'utilisateurs? Est-elle représentative de la population au sens large, compte tenu également d'autres groupes susceptibles d'être indirectement concernés?
- Avez-vous évalué si certaines personnes ou certains groupes pourraient subir de manière disproportionnée des effets négatifs?
- D'autres équipes ou groupes présentant différents parcours professionnels et expériences vous ont-ils fait parvenir des réactions?

Participation des parties prenantes:

✓ Avez-vous réfléchi à un mécanisme pour inclure la participation de différentes parties prenantes dans la mise au point et l'utilisation du système d'IA?

✓ Avez-vous préparé la voie à l'introduction du système d'IA au sein de votre organisation en informant et en mobilisant au préalable les travailleurs concernés et leurs représentants?

6. <u>Bien-être sociétal et environnemental</u>

IA durable et respectueuse de l'environnement:

✓ Avez-vous mis en place des mécanismes pour mesurer l'impact environnemental de la mise au point, du déploiement et de l'utilisation du système d'IA (par exemple, énergie consommée par les centres de données, type d'énergie consommée par les centres de données, etc.)?

✓ Avez-vous prévu des mesures pour réduire l'impact environnemental du cycle de vie de votre système d'IA?

Incidence sociale:

✓ Lorsque le système d'IA interagit directement avec des êtres humains:

- Avez-vous évalué si le système d'IA encourage les êtres humains à développer de l'attachement et de l'empathie pour le système?
- Avez-vous veillé à ce que le système d'IA indique clairement que son interaction sociale est simulée et qu'il n'a nullement la capacité de «comprendre» et de «ressentir»?

✓ Avez-vous veillé à ce que les incidences sociales du système d'IA soient bien comprises? Par exemple, vous êtes-vous demandé s'il existe un risque de perte d'emplois et de perte de compétences de la main-d'œuvre? Quelles mesures ont été prises pour contrer ces risques?

Société et démocratie:

✓ Avez-vous évalué l'incidence plus large de l'utilisation du système d'IA sur la société, au-delà de l'utilisateur (final) individuel, par exemple les parties prenantes susceptibles d'être indirectement concernées?

7. <u>Responsabilité</u>

Auditabilité:

- ✓ Avez-vous mis en place des mécanismes pour faciliter l'auditabilité du système par des acteurs internes et/ou indépendants, en veillant pas exemple à la traçabilité et à la journalisation des processus et des résultats du système d'IA?

Minimisation et documentation des incidences négatives:

- ✓ Avez-vous réalisé une analyse des risques ou de l'impact du système d'IA qui tienne compte de différentes parties prenantes qui sont directement et indirectement concernées?
- ✓ Avez-vous mis en place des cadres de formation et d'éducation pour définir des pratiques en matière de responsabilité?
  - Quels travailleurs ou branches de travailleurs sont concernés? Le sont-ils au-delà de la phase de mise au point?
  - Ces formations portent-elles également sur le cadre juridique potentiellement applicable au système d'IA?
  - Avez-vous envisagé la mise sur pied d'un «comité d'examen pour l'IA éthique» ou d'un mécanisme similaire pour discuter des pratiques globales en matière de responsabilité et d'éthique, y compris des zones grises potentiellement floues?
- ✓ Outre les initiatives ou cadres internes destinés à contrôler l'éthique et la responsabilité, des orientations externes ou des processus d'audit ont-ils également été mis en place?
- ✓ Existe-t-il des processus permettant aux tiers (par exemple, fournisseurs, consommateurs, distributeurs/vendeurs) ou aux travailleurs de signaler de possibles vulnérabilités, risques ou biais dans le système/l'application d'IA?

Documentation des arbitrages:

- ✓ Avez-vous mis sur pied un mécanisme permettant de recenser les intérêts et les valeurs pertinents concernés par le système d'IA et les éventuels arbitrages entre eux?
- ✓ Quel processus utilisez-vous pour prendre des décisions relatives à ces arbitrages? Avez-vous veillé à ce que les décisions d'arbitrage soient documentées?

Voies de recours:

- ✓ Avez-vous mis en place un ensemble approprié de mécanismes permettant un recours en cas de préjudice ou d'effet néfaste?
- ✓ Avez-vous mis en place des mécanismes pour fournir des informations aux utilisateurs (finaux)/tiers à propos des possibilités de recours?

Nous invitons l'ensemble des parties prenantes à tester la liste d'évaluation en pratique et à nous faire parvenir leurs réactions concernant son potentiel de mise en œuvre, son exhaustivité, sa pertinence à l'égard de l'application ou du domaine d'IA spécifique, et concernant tout chevauchement ou toute complémentarité avec des processus existants en matière de conformité ou d'évaluation. Sur la base de ces commentaires, une version révisée de la liste d'évaluation pour une IA digne de confiance sera proposée à la Commission début 2020

Orientations essentielles dérivées du chapitre III:

- ✓ Adopter une **liste d'évaluation** pour une IA digne de confiance au stade de la mise au point, du déploiement ou de l'utilisation de systèmes d'IA, et l'adapter au cas d'utilisation spécifique du système.
- ✓ Garder à l'esprit qu'une liste d'évaluation de cette nature **ne sera jamais exhaustive**. Il ne suffit pas de cocher des cases pour garantir une IA digne de confiance. Il convient de déterminer des exigences, d'évaluer des solutions et de garantir l'amélioration des résultats de manière continue tout au long du cycle de vie du système d'IA, et de mobiliser les parties prenantes.

C. EXEMPLES DE POSSIBILITES ET DE PREOCCUPATIONS MAJEURES SOULEVEES PAR L'IA

(121) Dans la section qui suit, nous présentons des exemples de mise au point et d'utilisation d'une IA qui devraient être encouragés, ainsi que des exemples de situations dans lesquelles la mise au point, le déploiement ou l'utilisation d'une IA peuvent nuire à nos valeurs et peuvent soulever des préoccupations spécifiques. Un équilibre doit être trouvé entre ce qui devrait et ce qui peut être fait avec l'IA. Il convient en outre de rester vigilant à ce qui doit être évité avec l'IA.

1. Exemples de possibilités offertes par une d'IA digne de confiance

(122) L'IA digne de confiance peut représenter un formidable potentiel pour contribuer à l'atténuation des problèmes urgents auxquels notre société est confrontée, tels que le vieillissement de la population, l'accroissement des inégalités sociales et la pollution de l'environnement. Ce potentiel est également reflété au niveau mondial, par exemple avec les objectifs de développement durable des Nations unies.[57] La section qui traite de la manière d'encourager une stratégie européenne dans le domaine de l'IA qui réponde à certains de ces problèmes.

a. Action pour le climat et infrastructures durables

(123) S'il est vrai que la lutte contre le changement climatique devrait être une priorité absolue pour les décideurs politiques du monde entier, la transformation numérique et une IA digne de confiance présentent un énorme potentiel pour réduire l'incidence humaine sur l'environnement et permettre une utilisation efficiente et efficace de l'énergie et des ressources naturelles[58]. Une IA digne de confiance peut par exemple être combinée aux mégadonnées pour détecter les besoins en énergie de manière plus efficace, ce qui donnerait lieu à des infrastructures et à une consommation énergétiques plus efficaces[59].

(124) Dans les secteurs tels que les transports publics, des systèmes d'IA appliqués à des systèmes de transport intelligents[60] peuvent être utilisés pour réduire le plus possible les files, optimiser l'itinéraire, aider les personnes souffrant de problèmes de vue à être plus indépendantes[61], optimiser les moteurs efficaces d'un point de vue énergétique et renforcer ainsi les efforts de décarbonation et réduire l'empreinte environnementale, pour une société plus verte. Aujourd'hui, à l'échelle mondiale, une personne meurt toutes

57 https://sustainabledevelopment.un.org/?menu=1300.

58 Un certain nombre de projets de l'UE visent à développer les réseaux intelligents et le stockage de l'énergie, qui peuvent potentiellement contribuer au succès d'une transition énergétique soutenue par les technologies numériques, y compris via des solutions fondées sur l'IA et d'autres solutions numériques. Pour compléter le travail de ces projets individuels, la Commission a lancé l'initiative BRIDGE, qui permet aux projets de réseaux intelligents et de stockage de l'énergie qui sont en cours dans le cadre d'Horizon 2020 d'établir une vision commune sur des questions transversales: https://www.h2020-bridge.eu/.

59 Voir par exemple le projet Encompass: http://www.encompass-project.eu/.

60 De nouvelles solutions fondés sur l'IA contribuent à préparer les villes à la mobilité de demain. Voir par exemple le projet financé par l'UE appelé Fabulos: https://fabulos.eu/.

61 Voir par exemple le projet PRO4VIP, qui fait partie de la stratégie européenne Vision 2020 pour combattre la cécité évitable, principalement due au vieillissement. La mobilité et l'orientation faisaient partie des domaines prioritaires du projet.

les 23 secondes dans un accident de voiture[62]. Les systèmes d'IA pourraient contribuer à réduire considérablement le nombre de victimes, par exemple grâce à une amélioration des temps de réaction et à un meilleur respect des règles[63].

b. Santé et bien-être

(125) Les technologies relevant de l'IA digne de confiance peuvent être utilisées – et le sont déjà – pour rendre les traitements plus intelligents et mieux ciblés, et contribuer à la prévention des maladies mortelles[64]. Les médecins et professionnels de la santé peuvent potentiellement réaliser un examen plus précis et détaillé de données de santé complexes relatives à un patient, avant même l'apparition d'une maladie, et administrer un traitement préventif sur mesure[65]. Dans le contexte du vieillissement de la population de l'Europe, l'IA et la robotique peuvent être des outils précieux pour assister les prestataires de soins et favoriser les soins prodigués aux personnes âgées[66], ainsi que pour surveiller en temps réel l'état de santé des patients, et donc sauver des vies[67].

(126) Une IA digne de confiance peut également être utile à une plus grande échelle. Par exemple, elle peut examiner et déterminer des tendances générales dans le secteur des soins de santé et des traitements[68], ce qui mène à une détection plus précoce des maladies, à un développement plus efficace des médicaments, à des traitements davantage ciblés[69] et, en fin de compte, à un plus grand nombre de vies sauvées.

c. Éducation de qualité et transformation numérique

(127) Les nouveaux changements technologiques, économiques et environnementaux impliquent que la société doit devenir plus proactive. Les pouvoirs publics, les leaders des secteurs concernés, les établissements d'enseignement et les syndicats ont la responsabilité de faire entrer les citoyens dans la nouvelle ère numérique en veillant à ce qu'ils disposent des compétences requises pour occuper les emplois de demain. Les technologies relevant de l'IA digne de confiance pourraient contribuer à améliorer la précision des prévisions relatives aux emplois et aux professions qui seront le plus perturbés par la technologie, aux nouveaux rôles qui

62 https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

63 Le projet européen UP-Drive vise par exemple à apporter des solutions aux problèmes signalés en matière de transport, par des contributions permettant l'automatisation et la collaboration progressives des véhicules entre eux, facilitant ainsi un système de transports plus sûr, plus inclusif et plus abordable. https://up-drive.eu/.

64 Voir par exemple le projet REVOLVER (Repeated Evolution of Cancer): https://www.healtheuropa.eu/personalised-cancer-treatment/87958/, ou le projet Murab qui réalise des biopsies plus précises, et qui vise à diagnostiquer plus rapidement le cancer et d'autres maladies: https://ec.europa.eu/digital-single-market/en/news/murab-eu-funded-project-success-story.

65 Voir par exemple le projet Live INCITE: www.karolinska.se/en/live-incite. Ce consortium d'acteurs du secteur de la santé incite ce secteur à mettre au point des solutions d'IA et d'autres solutions TIC intelligentes qui permettent des modifications du mode de vie dans le processus peropératoire. L'objectif concerne de nouvelles solutions innovantes de santé en ligne capables d'influencer les patients de façon personnalisée, afin qu'ils prennent les mesures nécessaires, tant avant qu'après la chirurgie, dans leur mode de vie pour optimiser le résultat des soins.

66 Le projet CARESSES financé par l'UE concerne des robots destinés aux soins aux personnes âgées, centrés sur leur sensibilité culturelle: ils adaptent leur manière d'agir et de parler pour correspondre à la culture et aux habitudes des personnes âgées qu'ils aident: http://caressesrobot.org/en/project/. Voir également l'application d'IA appelée Alfred, un assistant virtuel aidant les personnes âgées à rester actives: https://ec.europa.eu/digital-single-market/en/news/alfred-virtual-assistant-helping-older-people-stay-active. En outre, le projet EMPATTICS (EMpowering PAtients for a BeTTer Information and improvement of the Communication Systems) effectuera des recherches pour définir la manière dont les professionnels de la santé et les patients utilisent les technologies TIC, y compris les systèmes d'IA, pour planifier les interventions avec les patients et surveiller la progression de leur état physique et mental: www.empattics.eu.

67 Voir par exemple MyHealth Avatar (www.myhealthavatar.eu), qui propose une représentation numérique de l'état de santé d'un patient. Le projet de recherche a lancé une application et une plateforme en ligne qui collecte les informations concernant votre état de santé numérique à long terme, et vous y donne accès. Cela se présente sous la forme d'un compagnon de santé pour toute la vie («avatar»). MyHealthAvatar prédit également les risques d'accident vasculaire cérébral, de diabète, de maladie cardiovasculaire et d'hypertension.

68 Voir par exemple le projet ENRICHME (www.enrichme.eu), qui lutte contre la perte progressive des capacités cognitives au sein de la population vieillissante. Une plateforme intégrée d'assistance à l'autonomie à domicile (AAD) et un robot de service mobile pour un suivi et des interactions à long terme aideront les personnes âgées à rester plus longtemps actives et indépendantes.

69 Voir par exemple l'utilisation de l'IA par Sophia Genetics, qui tire parti de l'inférence statistique, de la reconnaissance de modèles et de l'apprentissage automatique pour optimiser la valeur de la génomique et des données d'imagerie médicale: https://www.sophiagenetics.com/home.html.

seront créés et aux compétences qui seront nécessaires. Elles pourraient aider les pouvoirs publics, les syndicats et les secteurs concernés à planifier la (re)qualification des travailleurs, ainsi qu'offrir aux citoyens craignant un licenciement une voie de développement dans un nouveau rôle.

(128) En outre, l'IA peut s'avérer être un excellent outil pour combattre les inégalités en matière d'éducation et créer des programmes de formation personnalisés et adaptables qui pourraient aider chacun à obtenir de nouvelles qualifications, aptitudes et compétences en fonction de ses propres capacités d'apprentissage[70]. Cela pourrait augmenter la vitesse d'apprentissage et améliorer la qualité de l'enseignement – de l'école primaire à l'université.

2. Exemples de préoccupations majeures soulevées par l'IA

(129) La violation de l'un des éléments constitutifs d'une IA digne de confiance est de nature à susciter une préoccupation majeure. Un nombre important des préoccupations présentées ci-dessous tiennent déjà aux exigences juridiques en vigueur qui, étant contraignantes, doivent par conséquent être respectées. Toutefois, même dans les cas où la conformité avec les exigences juridiques a été démontrée, celles-ci pourraient ne pas répondre à l'éventail complet de préoccupations éthiques susceptibles d'être soulevées. Étant donné que notre conception de l'adéquation des règles et principes éthiques est en constante évolution et peut changer au fil du temps, la liste non exhaustive suivante de préoccupations pourrait à l'avenir être raccourcie, élargie, modifiée ou actualisée.

a. Identifier et suivre des individus avec l'IA

(130) L'IA permet une identification encore plus efficace de personnes par des entités tant publiques que privées. La reconnaissance faciale et d'autres méthodes involontaires d'identification utilisant des données biométriques (à savoir, détecteur de mensonges, évaluation de la personnalité au moyen de micro-expressions et détection vocale automatique) sont des exemples notoires de technologies évolutives d'identification fondées sur l'IA. L'identification d'individus est parfois le résultat souhaitable, lorsqu'elle s'accompagne de principes éthiques (par exemple pour la détection de fraude, de blanchiment de capitaux, ou de financement du terrorisme). Toutefois, l'identification automatique soulève de fortes préoccupations de nature tant juridique qu'éthique, car elle peut avoir des effets inattendus à de nombreux niveaux psychologiques et socioculturels. Pour préserver l'autonomie des citoyens européens, il est nécessaire d'utiliser de manière proportionnée les techniques de contrôle dans le domaine de l'IA. Pour parvenir à mettre en œuvre une IA digne de confiance, il sera essentiel de définir clairement si, quand et comment l'IA peut être utilisée aux fins de l'identification automatique de personnes, ainsi que de faire la distinction entre l'identification d'un individu et le fait de le suivre à la trace, et entre une surveillance ciblée et une surveillance de masse. L'application de telles technologies doit être clairement justifiée dans la législation applicable[71]. Lorsque la base juridique pour une activité de cette nature est le «consentement», des moyens pratiques[72] doivent être développés pour permettre qu'un consentement éclairé et vérifié soit automatiquement recensé par une IA ou des technologies équivalentes. Cela s'applique également à l'utilisation de données à caractère personnel «anonymes» pouvant être re-personnalisées.

b. Systèmes d'IA cachés

[70] Voir par exemple le projet MaTHiSiS, visant à offrir une solution pour l'apprentissage fondé sur les affects dans un environnement d'apprentissage confortable, comprenant des dispositifs technologiques et des algorithmes de pointe: (http://mathisis-project.eu/). Voir également IBM Watson Classroom ou la plateforme Century Tech.

[71] Il convient à cet égard de rappeler l'article 6 du RGPD, qui prévoit notamment que le traitement de données n'est licite que s'il s'appuie sur une base juridique valable.

[72] Comme le montrent les mécanismes actuellement utilisés sur l'internet pour donner un consentement éclairé, les consommateurs accordent en général leur consentement sans véritable considération. Ces mécanismes ne peuvent dès lors que difficilement être qualifiés de pratiques.

(131) Les êtres humains devraient toujours savoir s'ils interagissent directement avec un autre être humain ou une machine, et les professionnels de l'IA ont la responsabilité de veiller à ce que ce soit effectivement le cas. Les professionnels de l'IA doivent par conséquent veiller à ce que les êtres humains soient informés du fait qu'ils interagissent avec un système d'IA – ou soient en mesure de le demander et de le confirmer – (par exemple, au moyen de clauses de non-responsabilité claires et transparentes). Il convient de noter que des cas limites existent et sont de nature à compliquer la question (par exemple, une voix filtrée par IA appartenant à un être humain). Il convient de garder à l'esprit que la confusion entre êtres humains et machines pourrait entraîner de multiples conséquences, telles qu'un attachement, une certaine influence, ou une réduction de la valeur accordée à la qualité d'être humain[73]. Le développement de robots humanoïdes[74] doit par conséquent faire l'objet d'une évaluation éthique minutieuse.

c. Notation des citoyens assistée par l'IA en violation des droits fondamentaux

(132) Les sociétés doivent s'efforcer de protéger la liberté et l'autonomie de tous les citoyens. Toute forme de notation des citoyens peut entraîner la perte de cette autonomie et mettre en péril le principe de non-discrimination. La notation ne doit être utilisée que si elle se justifie clairement et lorsque les mesures sont proportionnées et équitables. La notation normative de citoyens (évaluation générale de la «personnalité morale» ou de l'«intégrité éthique») dans *tous* les aspects et à grande échelle de la part des autorités publiques ou d'acteurs privés menace ces valeurs, notamment lorsqu'il y est recouru de manière non conforme aux droits fondamentaux et de manière disproportionnée sans objectif légitime délimité et communiqué.

(133) De nos jours, la notation des citoyens – à grande ou plus petite échelle – est déjà souvent utilisée dans le cadre de notations purement descriptives et spécifiques à un domaine (par exemple, systèmes scolaires, apprentissage en ligne et permis de conduire). Même dans ces applications plus étroites, une procédure totalement transparente doit être mise à la disposition des citoyens et comprendre des informations sur le processus, l'objectif et la méthodologie de la notation. Il convient de souligner que la transparence ne peut prévenir la discrimination ou garantir l'équité. Il ne s'agit en outre pas de la panacée face au problème de la notation. Idéalement, il devrait être possible de retirer sa participation au mécanisme de notation sans préjudice – dans le cas contraire, des mécanismes pour contester et rectifier les notes devraient être disponibles. Cela est particulièrement important dans les situations présentant des asymétries de pouvoir entre les parties. De telles options de retrait, qui sont nécessaires dans une société démocratique, doivent être garanties au stade de la conception de la technologie dans les cas où cela est nécessaire pour veiller au respect des droits fondamentaux.

d. Systèmes d'armes létales autonomes (SALA)

(134) À l'heure actuelle, un nombre inconnu de pays et d'industries recherchent et développent des systèmes d'armes létales autonomes, allant de missiles capables de sélectionner des cibles à des calculateurs autoadaptatifs dotés de compétences cognitives pour décider par qui, quand et où des combats peuvent être menés sans intervention humaine. Cela suscite des préoccupations éthiques fondamentales, comme le fait qu'une course à l'armement incontrôlable à un niveau jamais égalé dans l'histoire pourrait en résulter, ainsi que des contextes militaires dans lesquels le contrôle humain a presque totalement été abandonné et les risques de défaillances ne sont pas éliminés. Le Parlement européen a appelé à l'élaboration urgente d'une position commune contraignante pour régir les questions éthiques et juridiques du contrôle humain, de la surveillance, de la responsabilité et de la mise en œuvre du droit international relatif aux droits de l'homme, du droit humanitaire international et des stratégies militaires[75]. Rappelant l'objectif de l'Union européenne de

[73] Madary et Metzinger (2016). Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology. Frontiers in Robotics and AI, 3(3).

[74] Cela s'applique également aux avatars fondés sur l'IA.

[75] Résolution 2018/2752(RSP) du Parlement européen.

promouvoir la paix, tel que consacré à l'article 3 du traité sur l'Union européenne, nous saluons la résolution du Parlement du 12 septembre 2018 et tous les efforts y relatifs dans le domaine des SALA, que nous aspirons à soutenir.

e. Préoccupations potentielles à plus long terme

(135) La mise au point de l'IA reste spécifique à un domaine et requiert des scientifiques et ingénieurs humains correctement formés pour définir ses objectifs avec précision. Toutefois, en extrapolant à un horizon plus lointain, certaines grandes préoccupations à long terme peuvent faire l'objet d'hypothèses[76]. Une approche fondée sur les risques indique qu'il convient de continuer à tenir compte de ces préoccupations dans la perspective de possibles «inconnues inconnues» et «cygnes noirs»[77]. Les fortes incidences inhérentes à ces préoccupations, combinées à l'incertitude actuelle des évolutions correspondantes, requièrent des évaluations régulières de ces sujets.

D. CONCLUSION

(136) Le présent document constitue les lignes directrices en matière d'éthique dans le domaine de l'IA, élaborées par le groupe d'experts de haut niveau sur l'intelligence artificielle (GEHN IA).

(137) Nous reconnaissons les effets positifs que les systèmes d'IA ont déjà et continuerons à avoir, tant d'un point de vue commercial que pour la société. Nous sommes toutefois tout aussi soucieux de faire en sorte que les risques et autres effets néfastes auxquels sont associées ces technologies soient gérés de manière adéquate et proportionnée eu égard à l'application de l'IA concernée. L'IA est une technologie à la fois transformatrice et perturbatrice, et son évolution au cours des dernières années a été facilitée par la disponibilité de quantités gigantesques de données numériques, des avancées technologiques majeures en matière de puissance de calcul et de capacité de stockage, ainsi que par d'importantes innovations scientifiques et en ingénierie concernant les méthodes et outils de l'IA. L'incidence des systèmes d'IA sur la société et les citoyens se poursuivra sous des formes que nous ne pouvons pas encore imaginer.

(138) Dans ce contexte, il est important de mettre au point des systèmes d'IA dignes de confiance, car les êtres humains ne seront en mesure de tirer pleinement parti des avantages de l'IA en toute sérénité que s'ils peuvent se fier à la technologie, y compris aux processus et aux personnes qui la soutiennent. Lors de l'élaboration des présentes lignes directrices, notre ambition fondamentale a par conséquent consisté à tendre vers une IA digne de confiance.

(139) Une IA digne de confiance comporte trois éléments: 1) elle doit être licite, en assurant le respect des législations et réglementations applicables, 2) elle doit être éthique, en assurant l'adhésion à des principes et valeurs éthiques, et 3) elle doit être robuste, sur le plan tant technique que social, pour faire en sorte que, même avec de bonnes intentions, les systèmes d'IA ne causent pas de préjudices involontaires. Chaque élément est nécessaire mais ne suffit pas pour parvenir à une IA digne de confiance. L'idéal serait que ces trois éléments fonctionnent en harmonie et se chevauchent. Lorsque des tensions apparaissent, nous devrions nous efforcer de les aligner.

(140) Au chapitre I, nous avons articulé les droits fondamentaux et un ensemble de principes éthiques correspondants qui sont essentiels dans le contexte d'une IA. Au chapitre II, nous avons présenté sept exigences essentielles que doivent respecter les systèmes d'IA pour parvenir à une IA digne de confiance. Nous avons proposé des méthodes tant techniques que non techniques pouvant être appliquées aux fins de leur

[76] Si certains estiment que l'intelligence artificielle générale, la conscience artificielle, les agents moraux artificiels, la superintelligence ou l'IA transformatrice peuvent être des exemples de telles préoccupations à long terme (qui n'existent actuellement pas), celles-ci sont considérées par beaucoup comme irréalistes.

[77] Un événement «cygne noir» est un événement très rare dont l'impact est pourtant important – rare à tel point qu'il pourrait ne jamais avoir été observé. De ce fait, la probabilité qu'un tel événement survienne ne peut en général être estimée que de manière très incertaine.

mise en œuvre. Enfin, au chapitre III, nous avons proposé une liste d'évaluation pour une IA digne de confiance qui peut contribuer à concrétiser les sept exigences. Dans la section finale, nous avons présenté des exemples de potentiels bénéfiques et de préoccupations majeures soulevées par les systèmes d'IA, à propos desquels nous espérons encourager la poursuite des discussions.

(141) La perspective unique de l'Europe tient au fait qu'elle s'efforce de placer les citoyens au cœur de son action. Ce centrage sur les citoyens fait partie de l'ADN de l'Union européenne grâce aux traités sur lesquels elle s'est construite. Le présent document s'inscrit dans une vision qui encourage une IA digne de confiance qui, selon nous, doit être le fondement sur lequel l'Europe peut s'imposer comme leader du secteur des systèmes d'IA de pointe et innovants. Cette vision ambitieuse contribuera à garantir la prospérité des citoyens européens, sur le plan tant individuel que collectif. Notre but est de mettre en place une culture de l'«IA digne de confiance pour l'Europe», au moyen de laquelle chacun pourra récolter les fruits de l'IA dans le respect de nos valeurs fondatrices: les droits fondamentaux, la démocratie et l'état de droit.

GLOSSAIRE

(142) Le présent glossaire fait partie intégrante des lignes directrices et sert à comprendre les termes employés dans celles-ci.

Systèmes d'intelligence artificielle ou IA

(143) Les systèmes d'intelligence artificielle (IA) sont des systèmes logiciels (et éventuellement matériels) conçus par des êtres humains[78] et qui, ayant reçu un objectif complexe, agissent dans le monde réel ou numérique en percevant leur environnement par l'acquisition de données, en interprétant les données structurées ou non structurées collectées, en appliquant un raisonnement aux connaissances, ou en traitant les informations, dérivées de ces données et en décidant de la (des) meilleure(s) action(s) à prendre pour atteindre l'objectif donné. Les systèmes d'IA peuvent soit utiliser des règles symboliques ou apprendre un modèle numérique, et peuvent également adapter leur comportement en analysant la mesure dans laquelle l'environnement est affecté par leurs actions préalables.

(144) En tant que discipline scientifique, l'IA comprend plusieurs approches et techniques, telles que l'apprentissage automatique (dont l'apprentissage profond et l'apprentissage par renforcement sont des exemples spécifiques), le raisonnement automatique (qui comprend la planification, la programmation, la représentation des connaissances et le raisonnement, la recherche et l'optimisation) et la robotique (qui comprend le contrôle, la perception, les capteurs et les actionneurs, ainsi que l'intégration de toutes les autres techniques dans des systèmes cyberphysiques).

(145) Un document distinct élaboré par le GEHN IA et développant la définition des *systèmes d'IA* utilisée aux fins du présent document est publié parallèlement et s'intitule «Définition de l'IA: principales capacités et disciplines scientifiques».

Professionnels de l'IA

(146) On entend par professionnels de l'IA l'ensemble des personnes et organisations qui mettent au point (ce qui comprend la recherche, la conception et la fourniture de données), déploient (ce qui comprend la mise en œuvre) ou utilisent des systèmes d'IA, à l'exception de celles qui utilisent des systèmes d'IA en tant qu'utilisateurs finaux ou consommateurs.

Cycle de vie du système d'IA

(147) Le cycle de vie d'un système d'IA comprend sa phase de mise au point (dont la recherche, la conception, la fourniture de données et des essais limités), de déploiement (dont la mise en œuvre) et d'utilisation.

Auditabilité

(148) L'auditabilité désigne la capacité d'un système d'IA à faire l'objet d'une évaluation de ses algorithmes, de ses données et de ses processus de conception. Elle constitue une des sept exigences que doit respecter un système d'IA digne de confiance. Cela ne signifie pas nécessairement que les informations sur les modèles économiques et la propriété intellectuelle en lien avec le système d'IA doivent toujours être librement accessibles. Prévoir des mécanismes de traçabilité et de journalisation dès la phase de conception initiale du système d'IA peut contribuer à l'auditabilité du système.

Biais

(149) Le biais est une inclination au préjugé envers ou contre une personne, un objet ou un point de vue. Des biais peuvent se manifester de multiples manières dans les systèmes d'IA. Par exemple, dans les systèmes d'IA fondés sur les données, tels que ceux produits par apprentissage automatique, des biais présents dans la

[78] Les êtres humains conçoivent des systèmes d'IA directement, mais peuvent également avoir recours à des techniques d'IA pour optimiser leur conception.

collecte de données et l'entraînement peuvent être à l'origine de la présence de biais dans le système d'IA. Dans l'IA fondée sur la logique, comme les systèmes fondés sur des règles, le biais peut résulter de la manière dont un ingénieur des connaissances envisage les règles s'appliquant à un contexte particulier. Le biais peut également résulter de l'apprentissage en ligne et de l'adaptation par interaction. Il peut également se manifester à travers la personnalisation, par laquelle les utilisateurs reçoivent des recommandations ou des informations correspondant à leurs préférences. Il n'est pas nécessairement le résultat d'un biais humain et de la collecte de données par des êtres humains. Le biais peut, par exemple, se manifester dans les circonstances des contextes limités dans lesquels le système est utilisé, auquel cas il n'est pas possible de le généraliser à d'autres contextes. Un biais peut être positif ou négatif, intentionnel ou involontaire. Dans certains cas, le biais peut entraîner des résultats discriminatoires et/ou injustes, qualifiés de biais injustes dans le présent document.

Éthique

(150) L'éthique est une discipline qui est un sous-champ de la philosophie. De manière générale, elle traite de questions telles que «Qu'est-ce qu'une bonne action?», «Quelle est la valeur d'une vie humaine?», «Qu'est-ce que la justice?», ou «Qu'est-ce qu'une bonne vie?». L'éthique théorique comprend quatre principaux domaines de recherche: i) la méta-éthique, qui concerne principalement la signification et la référence d'un énoncé normatif, et la question de savoir comment leurs valeurs de vérité peuvent être déterminées (le cas échéant); ii) l'éthique normative, les moyens pratiques de déterminer une conduite morale, en examinant les normes applicables aux bonnes et mauvaises actions et en attribuant une valeur aux actions spécifiques; iii) l'éthique descriptive, visant une analyse empirique des comportements moraux et croyances morales des personnes, et; iv) l'éthique appliquée, qui concerne ce que nous sommes obligés de (ou autorisés à) faire dans une situation spécifique (souvent une première) ou dans un domaine particulier de possibilités d'action (souvent sans précédent). L'éthique appliquée traite de situations réelles, dans lesquelles des décisions doivent être prises dans des délais limités, et souvent avec peu de rationalité. L'éthique en matière d'IA est en général considérée comme un exemple de l'éthique appliquée et se concentre sur les questions normatives soulevées par la conception, la mise au point, la mise en œuvre et l'utilisation de l'IA.

(151) Dans les débats d'éthique, les termes «morale» et «éthique» sont souvent employés. Le terme «morale» renvoie aux modèles concrets et factuels de comportements, d'habitudes et de conventions qui peuvent s'observer au sein de cultures et de groupes ou auprès d'individus spécifiques à un moment donné. Le terme «éthique» désigne une appréciation évaluative de ces actions et comportements concrets d'un point de vue systématique et théorique.

IA éthique

(152) Dans le présent document, le terme «IA éthique» désigne la mise au point, le déploiement et l'utilisation d'une IA qui garantit une conformité avec les normes éthiques, y compris les droits fondamentaux en tant que droits moraux spéciaux, les principes éthiques et les valeurs essentielles qui s'y rapportent. Il s'agit du deuxième des trois éléments essentiels pour parvenir à une IA digne de confiance.

IA centrée sur l'humain

(153) L'approche de l'IA centrée sur l'humain s'efforce de garantir que les valeurs humaines soient un élément central de la mise au point, du déploiement, de l'utilisation et du contrôle des systèmes d'IA, en veillant au respect des droits fondamentaux, y compris ceux consacrés par les traités de l'Union européenne et la charte des droits fondamentaux de l'Union européenne, qui se rejoignent tous dans un fondement commun ancré dans le respect de la dignité humaine, en vertu duquel l'être humain jouit d'un statut moral unique et inaliénable. Cela implique également la prise en compte de l'environnement naturel et des autres êtres vivants qui font partie de l'écosystème humain, ainsi qu'une approche durable permettant l'épanouissement des générations à venir.

Méthode de l'équipe rouge («red teaming»

(154) La méthode de l'équipe rouge («red teaming») est une pratique par laquelle une «équipe rouge», c'est-à-dire un groupe indépendant, met au défi une organisation d'améliorer son efficacité en assumant un rôle ou un point de vue antagoniste. Cette pratique sert notamment à l'identification et à la résolution des failles potentielles en matière de sécurité.

Reproductibilité

(155) La reproductibilité est une indication de la mesure dans laquelle une expérience en matière d'IA, dans le cadre d'un essai, présente un comportement identique lorsqu'elle est répétée dans les mêmes conditions.

IA robuste

(156) La robustesse d'un système d'IA englobe tant sa robustesse technique (adaptation à un contexte donné, tel que le domaine d'application ou la phase du cycle de vie) que sa robustesse d'un point de vue social (le système d'IA tient dûment compte du contexte et de l'environnement dans lesquels il fonctionne). Cela est essentiel pour garantir que, même avec de bonnes intentions, aucun préjudice involontaire ne puisse survenir. La robustesse est le troisième des trois éléments nécessaires pour parvenir à une IA digne de confiance.

Parties prenantes

(157) On entend par parties prenantes tous ceux qui mettent au point, conçoivent, déploient, utilisent l'IA ou mènent des recherches dans le domaine, ainsi que ceux qui sont (directement ou indirectement) concernés par l'IA – y compris, mais sans s'y limiter, les entreprises, organisations, chercheurs, services publics, institutions, organisations de la société civile, pouvoirs publics, régulateurs, partenaires sociaux, particuliers, citoyens, travailleurs et consommateurs.

Traçabilité

(158) La traçabilité d'un système d'IA désigne la capacité de suivre les données du système et les processus de mise au point et de déploiement du système, en général au moyen d'une identification documentée.

Confiance

(159) Nous reprenons la définition suivante tirée de la littérature: «La confiance se définit comme: 1) un ensemble de convictions spécifiques portant sur la bienveillance, la compétence, l'intégrité et la prévisibilité (convictions en matière de confiance); 2) la volonté d'une partie de dépendre d'une autre partie dans une situation risquée (intention de faire confiance), ou 3) la combinaison de ces éléments.»[79] Si la «confiance» n'est en général pas une propriété que l'on prête aux machines, le présent document vise à souligner l'importance d'être capable de faire confiance non seulement dans le fait que les systèmes d'IA sont conformes sur le plan juridique, respectent l'éthique et sont robustes, mais aussi que cette confiance peut être accordée à l'ensemble des personnes et des processus impliqués dans le cycle de vie du système d'IA.

IA digne de confiance

(160) Une IA digne de confiance comporte trois éléments: 1) elle doit être licite, en assurant le respect des législations et réglementations applicables, 2) elle doit être éthique, en assurant le respect de principes et de valeurs éthiques, ainsi qu'en assurant l'adhésion à ces principes et valeurs, et 3) elle doit être robuste, sur le plan tant technique que social, pour faire en sorte que, même avec de bonnes intentions, les systèmes d'IA ne causent pas de préjudices involontaires. Une IA digne de confiance concerne non seulement la fiabilité du système d'IA en tant que tel, mais comprend également la fiabilité de l'ensemble des processus et acteurs qui font partie du cycle de vie du système.

Personnes et groupes vulnérables

[79] Siau, K., Wang, W. (2018), Building Trust in Artificial Intelligence, Machine Learning, and Robotics, *CUTTER BUSINESS TECHNOLOGY JOURNAL* (31), p. 47–53.

(161) Il n'existe pas de définition juridique communément ou largement admise de la notion de «personnes vulnérables», étant donné leur hétérogénéité. La raison pour laquelle une personne ou un groupe est vulnérable dépend souvent du contexte. Les événements temporaires de la vie (comme l'enfance ou la maladie), les facteurs de marché (comme l'asymétrie de l'information ou le pouvoir de marché), les facteurs économiques (comme la pauvreté), les facteurs identitaires (comme le sexe, la religion ou la culture) ou d'autres facteurs peuvent jouer un rôle. La charte des droits fondamentaux de l'Union européenne prévoit les motifs suivants à son article 21 sur la non-discrimination, qui peuvent constituer un point de référence parmi d'autres: le sexe, la race, la couleur, les origines ethniques ou sociales, les caractéristiques génétiques, la langue, la religion ou les convictions, les opinions politiques ou toute autre opinion, l'appartenance à une minorité nationale, la fortune, la naissance, un handicap, l'âge et l'orientation sexuelle. D'autres articles de la législation régissent les droits de groupes spécifiques, outre ceux énumérés ci-dessus. Toute liste de cette nature ne saurait être exhaustive et peut varier au fil du temps. Un groupe vulnérable est un groupe de personnes partageant une ou plusieurs caractéristiques de vulnérabilité.

Le présent document a été élaboré par les membres du groupe d'experts de haut niveau sur l'IA

énumérés ci-dessous par ordre alphabétique

Pekka Ala-Pietilä, président du GEHN IA
AI Finland, Huhtamaki, Sanoma
Wilhelm Bauer
Fraunhofer
Urs Bergmann – Co-rapporteur
Zalando

Mária Bieliková
Université technique de Slovaquie à Bratislava
Cecilia Bonefeld-Dahl – Co-rapporteure
DigitalEurope
Yann Bonnet
ANSSI
Loubna Bouarfa
OKRA
Stéphan Brunessaux
Airbus
Raja Chatila
Initiative de l'IEEE pour l'éthique des systèmes intelligents/autonomes et Université de la Sorbonne
Mark Coeckelbergh
Université de Vienne
Virginia Dignum – Co-rapporteure
Université d'Umeå
Luciano Floridi
Université d'Oxford
Jean-Francois Gagné – Co-rapporteur
Element AI
Chiara Giovannini
ANEC
Joanna Goodey
Agence des droits fondamentaux
Sami Haddadin
École de robotique et IA de Munich
Gry Hasselbalch
The thinkdotank DataEthics et Université de Copenhague
Fredrik Heintz
Université de Linköping
Fanny Hidvegi
Access Now
Eric Hilgendorf
Université de Würzburg
Klaus Höckner
Hilfsgemeinschaft der Blinden und Sehschwachen
Mari-Noëlle Jégo-Laveissière
Orange
Leo Kärkkäinen
Nokia Bell Labs
Sabine Theresia Köszegi
Université technique de Vienne
Robert Kroplewski
Avocat et conseiller du gouvernement polonais
Elisabeth Ling
RELX

Pierre Lucas
Orgalim – Europe's technology industries
Ieva Martinkenaite
Telenor
Thomas Metzinger – Co-rapporteur
Université Johannes Gutenberg de Mayence et Association des universités d'Europe
Catelijne Muller
ALLAI Netherlands et CESE
Markus Noga
SAP
Barry O'Sullivan, vice-président du GEHN IA
University College de Cork
Ursula Pachl
BEUC
Nicolas Petit – Co-rapporteur
Université de Liège
Christoph Peylo
Bosch

Iris Plöger
BDI (Fédération allemande de l'industrie)
Stefano Quintarelli
Garden Ventures
Andrea Renda
Collège d'Europe et CEPS
Francesca Rossi
IBM
Cristina San José
Fédération bancaire de l'Union européenne
George Sharkov
Digital SME Alliance
Philipp Slusallek
Centre allemand de recherche en IA (DFKI)
Françoise Soulié Fogelman
Consultante en IA
Saskia Steinacker – Co-rapporteure
Bayer
Jaan Tallinn
Ambient Sound Investment
Thierry Tingaud
STMicroelectronics
Jakob Uszkoreit
Google
Aimee Van Wynsberghe – Co-rapporteure
Université technique de Delft
Thiébaut Weber
CES
Cecile Wendling
AXA
Karen Yeung – Co-rapporteure
Université de Birmingham

Aimee Van Wynsberghe et Karen Yeung ont été rapporteurs pour le présent document.

Pekka Ala-Pietilä préside le GEHN IA. Barry O'Sullivan est vice-président et coordonne la deuxième contribution du GEHN IA. Nozha Boujemaa, vice-présidente jusqu'au 1er février 2019, chargée de la coordination de la première contribution, a également contribué au contenu du présent document.

Nathalie Smuha a fourni un soutien rédactionnel.